Configuration of Digital Repositories

Rushil Vallabh
VLLRUS002
University of Cape Town

ABSTRACT

For configuring a digital repository, it is important to understand what tasks are needed to be performed by the system, such as an administrative interface to configure access control or search and retrieve capabilities. Exploring previous work on digital repositories reference models can help adapt previous library strengths into a new system. When configuring the digital repository, the OAIS reference model can be used to model functions and implement an easy-to-use interface with separate functions for regular users and administrators for administrative tasks such as adding or removing content. Adapting a tool like digiflow can help reduce errors caused by poor workflow management in a repository.

Introduction

The project we set out to do is to improve the current user interface for administrators of the SimpleDL repository toolkit. Digital repositories are described as a way of storing and using digital objects [1]. It is a digital collection for content to be deposited such as academic work. Digital repositories offer services such as searching, retrieving , identifying, importing and exporting of its digital items [2]. The difference between digital repositories and the World-Wide Web is the difference in selection and organization. When a new digital object is added to the system it is integrated using its metadata without manual integration[10]. They have existed for years and play a crucial part in the importance of educational and historical institutions to support research, learning and administrative purposes. However, configuration and installation interfaces in earlier versions digital repositories were not a focus point for end users [20]. Administrators had to configure digital library tools themselves. This literature review serves the purpose of reviewing work done on generalized configuration settings for digital repositories as well as design of digital repositories to gain insight on the best methods to improve configuration tasks. Common repositories will be explored such as DSpace and Fedora etc. – flexible extensible digital and repository architecture.

Administrative configuration will be a focused topic. Administrative configuration is responsible for meeting standards being national and international for security and backup to be able to a trusted repository [3].

Tasks that are needed to be performed via configuration include allowing of retrieval of information. This should be possible through repository architecture and as well as open-source standards as it allows for storage of metadata [4]. Configurations are accessible through interfaces such as admin interfaces and have parameters such as control, setup, settings, and review.

In this paper, we will first present the current design of digital repositories and then discuss reference frameworks and models used for building digital repositories. The following sections will consist of a discussion of configuration in general and then configuration for digital libraries.

1 Design of digital repositories

1.1 Dspace

Dspace is designed to manage research data, such as publications and research materials, in a preservation-minded manner in a repository. The goal of the digital repository is to provide long-term access and visibility over time [5]. DSpace's purpose was to allow users to access research data with ease.

The interface used is a web-based interface that separates the user's screen, which focuses on searching and submitting, from the system administrator's screen. When a user clicks on an item, it opens in the user's browser with their chosen applicable format [5]. Additionally the DSpace web interface offers multiple home pages of community and collection which are configured by different communities. The home page offers searching functions with items containing specific metadata [6]. The Dublin core within each item can accessed however it requires configurations for authorizations. Title, language, and submission date are required fields when using the Dublin Core metadata standard to describe objects. Although not required, additional fields can be configured to be included [5]. However, it is important to note that DSpace does not have an advanced distribution mechanism but FEDORA can be used to address it [6].

The DSpace program offers a customizable workflow and submission process to be able to appeal to different useability for users [7]. Alternately, DSpace administrators can use import/export features like the command-line program Item Importer, which enables users to import collections of objects to the system. This is made possible by the directory structure of DSpace’s archive format used to store the system imports.

Exclusive named directories with all the information required to import an item can be found in the top archive directory. The Dublin Core file and another file providing the file names of the objects are the two files used by sub-folders. The content will be imported into the specific collection of the digital repository after the tool has run this. DSpace is suited for large organizations because of its customizable workflow [9].

1.2 Fedora

A huge digital repository, Fedora - Flexible Extensible Digital Object Repository Architecture focuses on a componentized framework that enables the creation of distributed digital repositories [7]. This framework can help build Digital library systems. A step up to DSpace, Fedora allows for dissemination as it has web services based on facilitating such tasks.

The Web Service Descriptive Language is used to explain the extensibility of Fedora's APIs [7]. The disseminations are used to show the various presentations an object can make. With additional focus on various sets of data, this can be used to exhibit Fedora objects in detail.

The design of the Fedora architecture is templates for data objects which includes digital resources, respective metadata and links to other software tools and services that are configured to bring the content is desired ways [8]. By utilizing services that are described by behavior objects, the Fedora digital library gives users access to data objects. To execute the system, runtime bindings and metadata about tool operations are stored in behavior objects. The WSDL is the Web Services Description Language.

The open source Fedora digital repository system makes use of open APIs that are made available as web services. The Web Service Definition Language – WSDL describes The Web Services Exposure Layer of the three layers of the Fedora system [8]. The management service for configuring repository administrative tasks, API has an open interface. These operations, which are implemented as a web service with SOAP support, include those required for users to generate and maintain digital goods.

1.3 Greenstone

The Greenstone repository is designed to work in a low-bandwidth environment through the use of a local Web server on a user's desktop. Greenstone uses an easy-to-create and configurable plugin to import digital items into the repository [9] It is designed for non-specialist users to compile digital objects into collections [11]. Importing documents undergoes a process of converting artifacts into a Greenstone Markup Language similar to an HTML format. The repository is set up so that for each collection there is a separate directory, and each directory contains five subdirectories: imports, archives, indexes, building directories, and supporting files [10]. The configuration files exist in the supporting files directory.

To identify objects within the system, the content is hashed to provide a unique identifier, or OID. This is to check that mirroring the data has no unintended consequences. After being imported, the items are saved in the corresponding subdirectories. [10]

Creating a new collection requires the creation of a collection configurable file. The information needed is the name of the directory and an email which users can use for support [10]. For searching, Greenstone enables the development of collections of non-textual content like video and audio. [11]. This repository is mainly suited for searching and indexing [9].

1.4 SimplyCT

SimplyCT was designed to work on low infrastructure and low cost of maintenance [9]. The repository delivers a Wed-based DRS done on work on an already known file hierarchy [12]. The system supports search and browse functionalities with an additional section for administrators to perform tasks such as control of an access option to the system.

The repository is based on design principles from past DLSes [13]. The DLSes configures itself depending on the data rather than having the user manually conform to the required structure and format, for example. There are no API structures needed as SimplyCT works on low infrastructure, so basic access to files is suitable. It is easy to preserve data when copied, as by copying the directory, preservation of data and services is ensured. There are no fixed services for searching, as searching can be done over the metadata and annotations. Instead of categorizing metadata, peer sub collections build layers of meaning on the digital objects as they are stored.

2. Models used for configuration of repositories: OAIS

The Open Archival Information System model is a reference model used for the modeling of the functions of a digital repository [14]. The framework can serve as an independent technology and community model that derives the components and systems needed to ensure the preservation and management of digital objects over a period of time. An information package, which is a sort of encapsulation for digital items, is how they move through the OAIS architecture [22]. The object is included in this package, along with data like preservation description information (PDI), which provides all the data required for storage to support search and retrieval capabilities of digital repository.

Archival storage, obtaining and storing archival information packages, placing them in permanent storage, and maintaining the storage hierarchy are the services that OAIS provides [15]. Moreover, OAIS offers a service that examines submissions to make sure they respect archive requirements and manage software and hardware setup. OAIS is primarily intended for use in repositories to preserve digital objects. The OAIS reference model has an advantage in that it has an abstract design, which allows for flexibility when configuring a repository with OAIS. [14]

3. Foundations of building digital repository

When building a digital library, it is important to note that it consists of many complex elements that arise from frameworks of different abstraction levels [16]. The following models/architecture show the building blocks to configure a better digital library. The reference model, which includes typical ideas and applications, is utilized to evaluate the digital libraries that are already available. To find abstract answers to the principles provided in the reference model, we employ the reference architecture. The system's standards and specifications are finalized using the concrete architecture. The Web Services Application Framework will be the run-time environment for this layer, for example.

To improve on the work that the reference model has gathered, there are requirements needed. The new system must have backwards compatibility, which allows the system to retain strengths from a previous design of a digital library [19]. Several levels of customization are needed while configuring the library's customization features as there are many different groups of people with various needs engaged in the library's construction. For long term management of this system, software modularity must exist. Software for page rendering and indexing can accomplish this. The configuration of the digital library should be dynamic so that interface problems may be fixed, and new services can be introduced as needed at run time. There are many other components to consider however if the design is dynamic then other functionalities can be added at a later stage.

4. Configuration of open file systems

Managing files on a desktop is facilitated by systems such as Ext4 and XFS. To meet the needs of a user, these systems are designed with many parameters that are configured and controlled by utilities. With many configurable parameters, issues may arise [17]. One of the issues that arise from configuration is configuration dependency, which means the different relations among the parameters of configurations.

When designing the configurations for the digital repository, it is important to deal with conflicts without affecting the end user, but doing so often leads to further unwanted conflicts [17]. Understanding where these dependencies are is necessary as to know as to know when to fix it. A multilevel configuration dependency is a dependency scheme such that a parameter is dependent on many configurations, not just one. ConfD is framework to identify multilevel configuration dependencies. Yet, having fewer settings is a better option because having too many parameters increases the possibility of vulnerabilities and misunderstandings [23]. As an alternative, replacing the software modularity design with the implementation of every utility function in the system would eliminate the multi-level configuration dependencies.

4.1 Methods of configuring file systems

There are a few methods for configuring file systems at four different stages, such as creation, which generates the system's first configurations by making use of the utility mkfs [17]. mounting a file system, where some configurations can be stipulated via mount. Utilities have the ability to change the configuration; this can be done by changing the metadata associated with the system online as well as offline.

A file system is configured in stages using a multi-stage configuration process [17]. This approach entails segmenting the configuration process into smaller tasks, each of which focuses on a different element of the system. This approach gives a more workable way to alter, for instance, access controls without changing other features of the file system configurations.

5. Configuration tools

Configurations tools are available via an interface to administrators of the system. The ways to classify if a configuration tool is based on it handles complexity and heterogeneity [18]. This heterogeneity in terms of a digital repository is differences in metadata schemas and other parameters for configuration. Ensuring the tasks for the heterogeneity work together is vital in ensuring the functioning of the system.

The same task of a configuration can be done by any choice of a tool, no one tool does one designated task [24]. Therefore, the use of many configuration tools can help build a digital repository however the need for a single tool consisting of all elements is needed still.

5.1 Workflow management

In terms of workflow the workflow management deals with composite changes such as a change on one maintenance channel it must be accounted for on another part of the system [18]. There are workflows such as those to handle adding new users, adding the same document to the system with a different name, and file naming conventions. Ineffective workflow management can lead to system errors and data loss [25]. For a workflow management tool to be effective, such as having automatic detection of workflow errors and managing the errors. Digiflow is a proposed tool for workflow management, it supports the development of digital repositories. This tool prevents the need to manage workflow tasks manually but rather provides a user with the tool necessary for carrying out a task which reduces of unintentional error.

6. Usability of configuration tools

For a configuration to be useable, a few features are needed to make up the configuration tool [18]. This being ease of the language, support for testing and monitoring infrastructure. The ease of language section forms part of administrative tasks. The language must be able to replace the already in place tools. The system configuration should be simple to use and not too complex. There should be support for testing such that the system uses mechanisms such as a dry-run mode to test the changes in the system first by replicating them in a virtual environment. A configuration tool must have its infrastructure monitored. This can either be a graphical system or a language system. The user can check the state of the system and data with the specifications of the configuration.

7. Digital library configuration

For configuring DSpace administrators do the task as it is too complex for some users as they would need training specifically for lower level components of configuration[20]. Lower-level elements include altering HTML code derived from JSP files. pertaining to the Deep Blue repository, a modified version of DSpace [21]. Lower level components include the item edit page, which makes it difficult to make additional changes to a single item because there are no confirmation messages indicating what has been done. Lower level components also include the remove option, which needs more secure access control, and the metadata editing task, which should have an option to save and exit or to exit and ignore any changes made.

In one concept of a digital library, the network of modules that made up the library communicated with one another using an instantiation of XML [19]. The parsed messages can be transformed using an XSL transform. Combining XML and XSL can offer a method for modifying the system's functionality while it is running without having to recompile the source code. This mechanism can be used to regulate a system's amount of reconfigurability.

Summary

Design

SimplyCT uses the lowest infrastructure and has the lowest cost of maintenance as compared to the other repositories. This repository uses a simple file hierarchy and DLses experiences to maintain a simple architecture. SimplyCT is mainly for smaller libraries. Greenstone is another repository designed to work with low resources such as low-bandwidth constraints. It focused on indexing and searching as opposed to SimplyCT which has more of a preservation aspect. This is accomplished by using easy-to-create and configurable plugin to import digital items into the repository and indexing is accomplished by the content being hashed to provide a unique identifier, or OID. DSpace is more suited for large organizations because of its customizable workflow. The main purpose of this repository was to allow for the digital artifacts to be accessible with ease, and the Dublin Core makes this possible. Fedora is a repository also suited for large organizations, Fedora makes use of a disseminator for digital objects, which is useful and a step up from DSpace as DSpace configuration does not feature dissemination.

Configuration

Systems like Ext4 and XFS, which have many configurable parameters managed by utilities, make it easier to manage files on desktops. Unfortunately, this could result in problems like configuration dependencies. Administrators, who must manage complexity and heterogeneity, have access to configuration tools via an interface and configure file systems using a multi-stage process. Lower-level components for configuring digital libraries like DSpace need more specific knowledge and care, and combining XML and XSL can provide a way to change the system's functioning without recompiling the source code. Less settings are preferable, and multi-level configuration dependencies would be removed if all system utilities were implemented.

For sizable digital repositories, configuration tools' scalability is crucial. Several workflows are used for managing tasks like adding new users and uploading processes, which deal with composite modifications. The deployment architecture of a system configuration tool begins with a centralized specification and consists of a translation agent and push/pull distribution. Language simplicity, support for testing, and monitoring infrastructure to compare the state of the system and data with the configuration's specifications are features that are required for an effective configuration tool. Also, using it shouldn't be too difficult.

Conclusion

In conclusion, for building an interface for administrative configuration tasks in a digital repository, there are many designs and reference models for digital repositories to be able to do so. Using these designs and models depends on the needs of the system and what configuration tools and properties are best suited.

References

Devendra S. Gobbur, (2007) Digital repositories: concepts and issues. https://core.ac.uk/reader/333967000 [1]

Krishnan Murugan. (2016). Digital Repositories: An Overview. https://www.researchgate.net/publication/308034171_Digital_Repositories_An_Overview [2]

Research Libraries Group (2002) Trusted Digital Repositories: Attributes and Responsibilities An RLG-OCLC Report

https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf [3]

Mandal S (2018). Designing and Configuration of Institutional Digital Repository through OPUS: An Integrated Open Access Framework. Inter. J. Acad. Lib.Info.Sci.6(1):8https://www.researchgate.net/publication/334731531_Desig ning_and_Configuration_of_Institutional_Digital_Repository_through_OPUS_An_Integrated_Open_Access_Framework [4]

Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., Tansley, R. and Walker, J.H., 2003. DSpace: An open source dynamic digital repository. https://dspace.mit.edu/bitstream/handle/1721.1/29465/DLib%20article%20January%202003.htm?sequence=1&isAllowed=y [5]

R. Tansley M. Bass D. Stuve M. Branschofsky D. Chudnov G. McClellan, "The DSpace institutional digital repository system: current functionality," 2003 Joint Conference on Digital Libraries, 2003. Proceedings., Houston, TX, USA, 2003, pp. 87-97,doi:10.1109/JCDL.2003.1204846.
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1204846 [6]

Douglas Gorton, 2007. PRACTICAL DIGITAL LIBRARY GENERATION INTO DSPACE WITH THE 5S FRAMEWORK: A Thesis Presented to the Faculty of
The Virginia Polytechnic Institute and State University In Partial Fulfillment of the Requirements of the Degree of Master of Science in Computer Science and Applications
https://vtechworks.lib.vt.edu/bitstream/handle/10919/31914/dgorton_thesis_final.pdf?sequence=1&isAllowed=y [7]

Thornton Staples, Ross Wayland, Sandra Payette. 2003. The Fedora Project.
An Open-source Digital Object Repository Management System. D-Lib Magazine Volume9Number4ISSN1082-9873
https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a17a7fc55fb142180dff749958db9a58170c5257 [8]

Hussein Suleman.2019. Reflections on Design Principles for a Digital Repository in a Low Resource Environment
https://pubs.cs.uct.ac.za/id/eprint/1331/1/ho_2019_lowresource.pdf [9]

Witten, I.H., Boddie, S.J., Bainbridge, D. and McNab, R.J., 2000, June. Greenstone: a comprehensive open-source digital library software system. In Proceedings of the fifth ACM conference on Digital libraries (pp. 113-121). https://dl.acm.org/doi/pdf/10.1145/336597.336650 [10]

Ravikumar, M.N. Ramanan, T., 2014. Comparison of greenstone digital library and DSpace: Experiences from digital library initiatives at eastern university, Sri Lanka. Journal of University Librarians Association of Sri Lanka, 18(2), pp.76-90. https://pdfs.semanticscholar.org/c067/0747a65c9adad3dd341b99a0feb6edf6bf0d.pdf [11]

Stuart Hammar Miles Robinson., SimplyCT. A Web-based Digital Repository System.https://projects.cs.uct.ac.za/honsproj/cgibin/view/2011/hammar_robinson.zip/Website/download/proposal/Project_Proposal.pdf [12]

Hoosein Suleman, 2012. The design and architecture of digital libraries. https://pubs.cs.uct.ac.za/id/eprint/828/1/iai_2011_architecture.pdf [13]

Gareth Knight, Mark Hedges,SHERPA DP Project, AHDS Executive (2008). Modelling OAIS compliance for disaggregated preservation services. International Journal of Digital Curation, 2(1). http://www.ijdc.net/article/view/25/14 [14]

Sibsankar Jana Mrinal Kanti Mondal Ujjal Marjit. (2009) Digital Preservation with Special Reference to the Open Archival Information System (OAIS) Reference Model:AnOverview.
https://ir.inflibnet.ac.in:8443/ir/bitstream/1944/993/1/3.pdf [15]

Leonardo Candela, Donatella Castelli, Pasquale Pagano, Constantino Thanos Yannis Ioannidis, Georgia Koutrika Seamus Ross Hans-Jörg Schek Heiko Schuldt (2007). Setting the foundations of digital libraries. D-Lib, 13(3/4). https://www.researchgate.net/profile/Leonardo-Candela/publication/225035738_Setting_the_Foundations_of_Digital_Libraries_The_DELOS_Manifesto/links/09e415076d4afe1f25000000/Setting-the-Foundations-of-Digital-Libraries-The-DELOS-Manifesto.pdf [16]

Tabassum Mahmud, Om Rameshwar Gatla, Duo Zhang, Carson Love, Ryan Bumann, and Mai Zheng (2019) {ConfD}: Analyzing Configuration Dependencies of File Systems for Fun and Profit. In 21st USENIX Conference on File and Storage Technologies (FAST 23) (pp. 199-214). https://www.usenix.org/system/files/fast23-mahmud.pdf [17]

Thomas Delaet Wouter Joosen Bart Vanbrabant (2010). A survey of system configuration tools. In Proceedings of the 23rd Large Installations Systems Administration (LISA) conference (pp. 1-14). Usenix association. https://www.usenix.org/legacy/event/lisa10/tech/full_papers/Delaet.pdf [18]

David Bainbridge, Katherine J. Don, George R. Buchanan, Ian H. Witten Steve Jones, Matt Jones, and Malcolm I. Barr (2004). Dynamic digital library construction and configuration. In Research and Advanced Technology for Digital Libraries: 8th European Conference, ECDL 2004, Bath, UK, September 12-17, 2004. Proceedings 8 (pp. 1-13). Springer Berlin Heidelberg. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a9f78d7c20e49cead4e58e7b16d724a5f62b28fa [19]

Nils Ko¨rber, Hussein Suleman (2008). Usability of digital repository software: A study of DSpace installation and configuration. In Digital Libraries: Universal and Ubiquitous Access to Information: 11th International Conference on Asian Digital Libraries, ICADL 2008, Bali, Indonesia, December 2-5, 2008. Proceedings 11 (pp. 31-40). Springer Berlin Heidelberg.
https://pubs.cs.uct.ac.za/id/eprint/492/1/2008_06_15_icadl_submission.pdf [20]

Ottaviani, Jim (2006). Deep Blue/DSpace Usability Summary (incl. School of Information Class Project).
https://deepblue.lib.umich.edu/bitstream/handle/2027.42/107021/Deep_Blue%28DSpace%29_usability_summary.pdf?sequence=1&isAllowed=y [21]

Alex Ball (2006) Briefing Paper: the OAIS Reference Model
http://www.ukoln.ac.uk/projects/grand-challenge/papers/oaisBriefing.pdf [22]

Tabassum Mahmud, Duo Zhang, Om Rameshwar Gatla, Mai Zheng (2022, June). Understanding configuration dependencies of file systems. In Proceedings of the 14th ACM Workshop on Hot Topics in Storage and File Systems (pp. 1-8). https://dl.acm.org/doi/pdf/10.1145/3538643.3539756 [23]

Paul Anderson. Edmund Smith (2005, December). Configuration Tools: Working Together. In LISA (pp. 31-37).
https://www.usenix.org/legacy/event/lisa05/tech/full_papers/anderson/anderson_html/ [24]

Ángeles S. Places, Antonio Fariña, Miguel R. Luaces, Óscar Pedreira, Diego Seco (2016) A workflow management system to feed digital libraries: proposal and case study. Multimed Tools Appl 75, 3843–3877 https://doi.org/10.1007/s11042-014-2155-3 [25]