Interoperability and Metadata Capability of Research Data Management's Tools: A Comparative Study of DSpace and EPrints

Dhanwantari Prakash Tripathi
Assistant Librarian,
Central Library, NIT Rourkela, Odisha, India
E-Mail: tripathidp@nitrkl.ac.in

Abstract:

With the emergence of multidisciplinary subjects and research in different domains, managing information and research data has become difficult task and important concern for higher educational universities and research institutions. To handle this challenging task, several data management platforms have also emerged. This paper mainly focusses on current practices the workflow of managing research data and identify the different stakeholders and further comparing two-well-known open source repository platform DSpace and E-Prints considering their architecture, metadata, different search features & mechanism and community support with acceptance. Metadata in any repository plays an important role in ensuring to find out the data and reuse by researchers / educational institutes even outside of the original research group. To complete this paper, the requirement of different stakeholders have also been taken into consideration to evaluate the features.

Keywords: Research Data Management, Open Source, DSpace, EPrints

1. Introduction:

These days huge number of research papers are published and awareness has increased among community about the data's complexity and diversity with its importance generated during research. The importance aspect and challenge is to manage these assets for information managers and researchers. They are mainly responsible for streamlining the intellectual output and scholarly communication at the time of preserving the data of research contributors and ensuring the proper licensing of their data / content. There has been gradual increase in research publications having open access policies and this has created opportunity for the development of many open source data management platforms. Academic and higher educational institutes have their own policies managing the research data and demand for structured infrastructure for supporting these activities with ease. Many alternatives are available but it is difficult for educational institutes/universities to go ahead by selecting the appropriate data management platform. Previous researchers have done many studies in this regard and carried out research covering different aspects of

implementations. Many research publications have given the access to their research data and found citation rates higher in comparison to those publications that don't give access. In research workflow, the data management is very important part and to have the proper management, the open source communities have found solutions for managing research data. Most of the repositories have the main concern of longtime preservation of research data but all of them having different requirements with different research domains. In most of the repositories, there is structural diversity in data with different metadata requirements to represent these records. The basic purpose of this paper is to evaluate and present two well-known open source data management platforms – **DSpace** and **EPrints** – and different features such as interoperability, architecture, capabilities of metadata handling, dissemination of data and information, search features and different search strategy with mechanism. Interoperability and metadata in particular play an important role in ensuring to find out the data and reuse by researchers / educational institutes even outside of the original research group.

2. Management of Research Data:

Management of research data is mainly concerned with proper organization of data. In research cycle, it starts from data entry then dissemination and archiving of research output with aim to make the research efficient and meet the expectations and requirements of the Institutes / Organizations. Data management mainly is concerned with creation of data, its plan with use, organization of data, proper structure and giving it suitable name. Further, it is important to make it to secure with access, store and back it up, find information resources, publish and is cited. Research data management has become a very common practice by research institutions to ensure that research outputs are organized and preserved overtime and properly. At present many research teams have adopted different known platforms to organize their research data properly and to further share with research community ensuring the preservation guarantee. Many stakeholders such as research scholars, educational institutions, metadata harvesters, metadata curators and developers are involved in description of dataset and workflow of data management for dissemination. These stakeholders play a major role in data repository's requirement for research data / output. Along with stakeholders, institutions are also encouraged to have their research data organized and preserved. These institutions value the compliance with standards of metadata to prepare them for including in networked environments for enhancing the visibility. The role of curator is to make sure the data quality and integrity is up to the mark for preservation. Generally, these curators are not an expert in research domain so they work very closely with research scholars for producing the quality metadata records. Long-term preservation of data mainly depends on

information / data specialists such as librarian or archivist including the expert involved in description of data. Data dissemination depends on harvesters, which index the content of repository. Here, Open Access Initiative for Metadata Harvesting (OAI-PMH) can play a good role for retrieving the metadata from other sources. It can also help to develop an interface to display data and resources properly indexed. However, choosing the right platform for any institute is never easy so it can be good starting point to compare the features of each platform to make a logical choice.

3. Platform Evaluation

Evaluation started with the selection of two major and well-known open source data management platforms operational at National Institute of Technology Rourkela namely **DSpace** and **Eprints**. Apart from these, there are many open source data repositories software, those may also be considered while selecting the platform for data management. However, DSpace and EPrints have capability of long-term preservation of data with different features including the different international standards such as OAI-PMH, OAIS, PREMIS and METS etc. Repositories platforms can be assessed with the following categories based on identified stakeholders.

a. Architecture of selected platform for preserving the data
b. Different features with challenges in implementation
c. Metadata Schema – Its standard, submission process and relevance to data
d. Capabilities of dissemination of information and data retrieval
e. Community support and also institutions adaptation

3.1. Architecture of the Repository:

While considering the architecture of repository's platform, every aspect must be given proper attention and must be selected considering all the aspects of Institute and further, selected data management platform must be deployed for testing purpose in any local machine first. This process will enable the institute to get ready for:

a. Signing the agreement with vendors or
b. Institution and customization of its own repository
c. Supporting the physical infrastructure including labor with maintenance costs.

Having support and contract with service provider by selected company may support the institute undoubtedly based on maintenance cost (monthly / annual). However, this approach may not be suitable to many institutions and they may not sign the agreement for sharing their research work / data in vendor's platform because this may bring situation of theft or loss of important and sensitive research data.

In this situation, it is always better to develop and have own platform with own infrastructure. DSpace and Eprints in this regard offers a better control to the institutions over the deposited research data. Because, these software can be installed and taken care by the research institutes itself. Both software have strong community and forum support that contribute and provide regular updates and additional plugins & extensions to meet specific needs of the research institutes. Both can be customized easily for better-improved interfaces with the help of extensions and plugins available freely or even paid also. DSpace and EPrints – supports and capable to create environment where teams can collaborate and manage the deposited resources. The data security even can be controlled by enabling the embargo periods and made available to the outside community after the embargo period.

3.1.1. DSpace – Open Source Digital Data Management Software

DSpace is one of the internationally accepted digital object / data management open source software. HP and MIT as joint project originally developed it in the year 2002. This software enables to establish institutional repository. The current stable release of DSpace is 6.3 and available to download with source-code from www.dspace.org. DSpace has been adopted and accepted worldwide by educational institutions, research community, and has capability to deal with management of research data and publications.

3.1.2. Architecture

DSpace functions based on three layers namely (a) Storage Layer (b) Business Layer and (c) Application Layer. The storage layer manages the physical storage of content and its metadata. The business layer manages the content of the archive, different e-people (users), and authorization of item/collection/community and workflow of DSpace. The application layer is responsible for containing the components that communicates with the world. Web user interface and the Open Archives Initiative protocol for metadata harvesting service is the best example of the same.

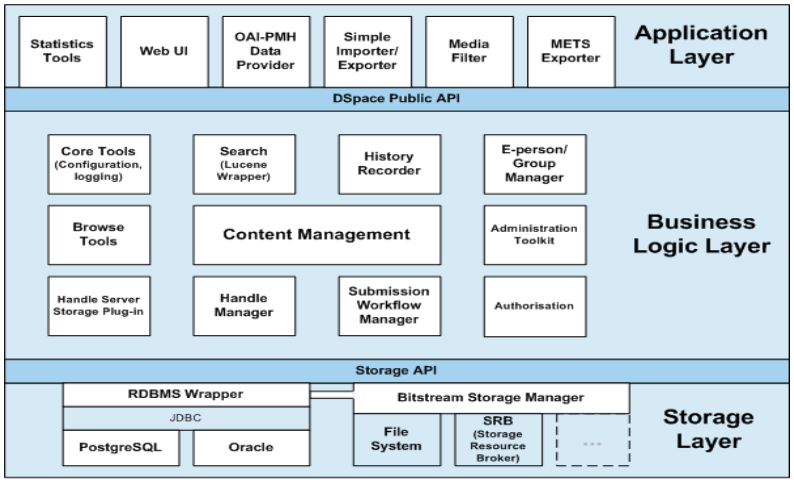

Image Source: https://wiki.duraspace.org/display/DSDOC6x/Architecture

3.1.3. Installation and Administration:

The installation of DSpace is possible on cross platforms such as Windows and Linux and is straightforward in comparison to other repository software. However, the customization to advance level may not be an easy task and may require the assistance of third party / external vendor. However, the developers are putting their efforts to make the software more extensible and flexible.

3.1.4. Metadata:

DSpace is supported with international metadata standard qualified Dublin Core by default and is capable to export data in many formats such as METS, MODS, RDF, MARC etc. However, it also supports the creation of custom metadata schema using XML. It also has the feature for using a controlled vocabulary with look-up option in submission form.

3.1.5. Interoperability:

DSpace also supports interoperability and is oriented towards open standards and protocols. It not only supports OAI-PMH and SWORD protocols but can also ingest and export information through other protocols too. DSpace also has SHERPA/RoMEO API look-up functionality but limited to authority control. DSpace is capable of integrating into other system and building a user interface in lightweight web application framework.

3.1.6. Content Management – Embargo and Preservation

DSpace has a good feature of creating embargoes (access restriction) on community/collection/items. DSpace is also capable of managing the digital objects and preserving them in proper way. However, while maintaining the repository, few institutes also keep separate server for preserving the deposited data. Archivematica in this regard may be of great help.

3.1.7. Statistics

DSpace has provision of Solr, which helps to logs internal events inside DSpace such as bitstream downloads, and workflows statistics. DSpace has also the option to display this information using Elasticsearch. In addition to this, Google Analytics may be used to store download and page views. The Edinburgh Research Archive in this regard has developed a Google Analytics integration module but this feature is available only JSPUI interface only that can track and display usage statistics at all level such as item collection and community level.

3.1.8. File Formats and Batch Importing

The best part of DSpace is that it supports all file types generated through computer. It may be text, pdf, video, audio, etc. The most challenging part may be while importing large research

datasets but the developers have given solutions as Globus and SWORD with workflow for batch importing through XML.

3.1.9. User Interface

DSpace has two interfaces – JSPUI and XMLUI and both are functional and fully responsive, easily customizable with third party support.

3.1.10. Search

DSpace uses a popular and robust search engine namely 'Lucene' that is open source in nature. The challenges may be in configuration and may require the support of programmer in some cases.

3.1.11. Support

There are many platforms where support can be obtained such as mailing lists, IRC channel, forum and third parties (vendors).

3.2.1. EPrints

Researchers of the University of Southampton, School of Electronics and Computer Science developed a free and open source software in 2000 and named it as 'EPrints'. It was designed for research papers, theses and teaching materials' management and archival but EPrints can accommodate any type of content. It can be downloaded from www.eprints.org

3.2.2. Architecture

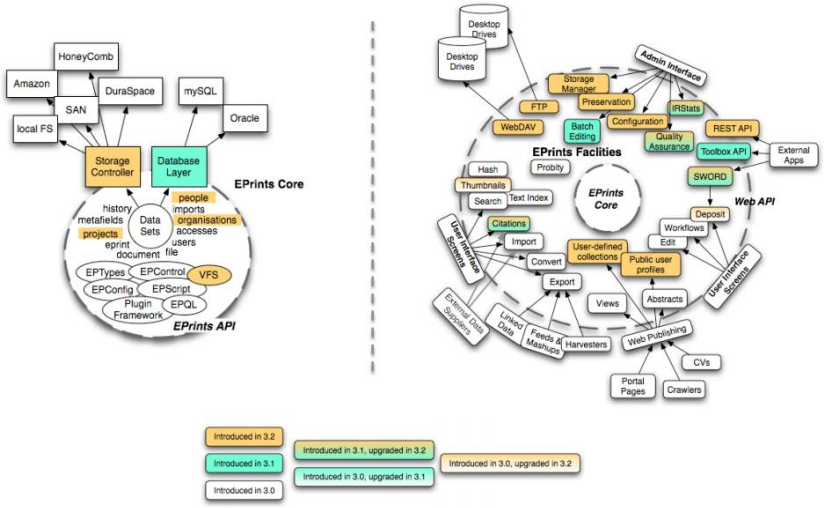

Image Source: http://files.eprints.org/712/

3.2.3. Installation / Administration

Similar to DSpace, EPrints software also can be installed following "turn-key" approach. The process of installation is very straightforward. It also has the administrative back-end that enables the configuration of EPrints at internal level. Additional plug-ins are also available at EPrints Bazaar to enhance its different services.

3.2.4. Metadata

EPrints supports controlled vocabulary and has provision of authority list that helps to manage high quality metadata. By default, EPrints has support for internal metadata standard Dublin Core and it can export to many numbers of formats such as METS, MODS etc.

3.2.5. Interoperability

EPrints supports OAI-PMH and SWORD and fully interoperable. Using XML, metadata of repository and directory structure can be exported but in some cases knowledge of advance level scripting is necessary and required.

3.2.6. Content Management – Embargo and Preservation

EPrints enables the submitter to restrict the access of content by embargo feature that restrict the content date wise. It also supports metadata and repository preservation and allows downloading all the files and metadata.

3.2.7. Statistics

EPrints supports a good feature namely IRStats package that allows keeping track of download counts of full-text documents. The statistics generated through IRStats can be viewed as table and graph.

3.2.8. File Formats and Batch Importing

EPrints accepts any kind of file in any format such as text, audio, video, etc. However, to support research datasets, customization is required in order to extent EPrints. Compared to other repository software, EPrints requires additional knowledge of Perl scripting for batch importing.

3.2.9. User Interface

EPrints has a user-friendly interface that allows the submitter to submit and manage files easily. The best part of EPrints is that it allows easy modification of workflow.

3.2.10. Search

EPrints supports Xapian search engine that allows search of all metadata fields, sort result by date of issue, name of the author, title and supports Boolean operator for advance level searching. PDF, Word and HTML is enabled with full-text indexing.

3.2.11. Support

EPrints supports training, documentation, support and customization as per requirement. There is mailing list namely EPrints-tech helps the EPrints user to get the assistance against problem faced while using the software.

Table: 1

Comparison of Evaluated and Selected Research Data Management Platforms

Criteria	Key Feature	DSpace	EPrints
Architecture of Platoform	Installation	Package Installation	Package Installation
	Location of Storage	Remote & Local (both)	Remote & Local (both)
	Cost Involvement	Manage Infrastructure	Manage Infrastructure
	Open Source	Available with Source Code	Available with Source Code
	Platform Customization	YES but Expertise Required	YES, Easily Customizable
	Embargo Period	YES, Access Restriction	YES, Access Restriction
	Open Architecture	Straightforward	Flexible Plugin Architecture
	DOI/Handle No.	YES	YES
Metadata	Required Fields	Title, Date of Issue	Title, Creators, Publication Details, Publication Titles
	Exporting Schemas	Any pre-loaded Schema	YES
	Schema Flexibilities	Flexible	Flexible
	Validation	YES	YES
Dissemination	API	YES	YES
	OAI-PMH Compliance	YES	YES
	Faceted Search	YES	YES
	Metadata Included	YES	YES
Others	Platform	Windows/Linux	Windows/Linux
	RDBMS	PostgreSQL	MySQL
	Network Environment	Supports	Supports
	Metadata Formats	DC, QDC, METS	DC, QDC, METS

3.3. Metadata for Preservation of Data

Metadata is a preservation key in order to manage the research data and output and can benefit metadata to production. DSpace and Eprints have ability to use different schemas that system administrator can configure and customize as per requirements. Both software are capable of exporting records in internationally accepted metadata schemas (MARC-XML and Dublin Core). However, DSpace is having another robust feature of exporting Dissemination Information Package (DIP) that includes records of metadata to enable the ingestion of exported records into long-term preservation workflow. DSpace natively supports different stages of data validation by

curators and researcher and enable the curation of data and metadata structure. In the process of data management, to track the changes in content is also an important issue and well taken care by DSpace and EPrints.

3.4. Interoperability and Dissemination of Data

Showing contents repository on other research platforms may definitely improve the visibility of data and reuse for further research work. All the well evaluated data management platforms permits the tools development and allow the external clients to expose metadata records with APIs to outside academic community but there may be some cases where difference will arise in standards compliance. DSpace and E-Prints – both supports the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). OAI-PMH is internationally accepted and used worldwide and is responsible for promoting interoperability between repositories and streamlining dissemination of data. It helps the metadata harvesters and index the repository content. Both the platforms supports and have capability of search the text freely, indexing the records for retrieving the data easily. These platforms have the feature of 'advanced' search that allows users to get specific information with restricted and specific keywords. Researchers may find information easily and pin pointed specific result from relevant and specific domains / collection / categories.

4. Selection of Repository Platforms:

The repository platform depends on developer's community to build, maintain and enhance the platform features. Many studies narrate the impact and comprehensiveness and National Institute of Technology Rourkela is one of them that maintains three different repositories using DSpace and EPrints and have been made available to the outside academic community. In not only National Institute of Technology Rourkela, but DSpace and EPrints have better acceptance by the academic community and stakeholders due to its robust features for management of research data.

Table: 2

Advantages of the Data Management Platforms (DSpace and EPrints)

Platforms	Key Advantages
DSpace	- Supports internationally accepted and comply with metadata schemas - Has a wide and strong open source community and forum - Has depth documentation prepared by developer community - Can be installed and managed by Institution itself - Represent structured metadata - Compatible with International protocol OAI-PMH
EPrints	- Repository with specification of high quality institutional open access collection

	- Saves time while depositing the items to repository, - Capable of importing metadata from other repositories, - Supports Internationally accepted metadata schemas, - Compatible with International Standards and protocols, - Supports and flexible with additional plugin to enhance services.

5. Concluding remarks:

Comparative study and evaluation describes that it is difficult for any institutes to select a platform without testing the performance, studying carefully the different features and requirement of all stakeholders. As a result, both – DSpace and EPrints – have capabilities of dealing with research data effectively but DSpace is more robust that allows to be updated and customized keeping the core functionalities intact. Many government institutions / organization for disclosing their data, to manage research data, are using DSpace. Curators as identified stakeholder may also favor DSpace as it enables DSpace System Administrators to create the parameter additional metadata schemas to be used for describing the resources. Researchers may also favor DSpace for easy deposit with easy data citation. Control over the stored data is an important factor and need to be considered. Many institutions may not wish to store their data out of their control and may like to implement a solution where research data can be stored in server completely under their control.

In the broader sense, DSpace is recommended as a good and robust platform to be installed in server instead of relying on external storage provided by external service providers. It may support better conditions for long-term preservation of research data.

References:

1. Amorim, R.C., Castro, J.A., da Silva, J.R., Ribeiro, C. (2015). A comparative study of platforms for research data management: interoperability, metadata capabilities and integration potential. In: Rocha, A., Costanzo, A. M. & Reis, L. P. (eds.) New contributions in information systems and technologies, vol. 1, pp. 101–112. Springer, Heidelberg
2. Corti, L., Eynden, V.d., Bishop, L., Woollard, M. (2014). Managing and Sharing Research Data: A Guide to Good Practice. SAGE Publications
3. Fay, E. (2010). Repository software comparison: building digital library infrastructure at LSE. Ariadne (2009), 1–11
4. Green, A., Macdonald, S., & Rice, R. (2009). Policy-making for Research Data in Repositories: A Guide. DISC-UK, Edinburgh
5. Lynch, C.A. (2003). Institutional repositories: essential infrastructure for scholarship in the digital age. Portal: Library and the Academy.

6. M.R. Barton and M. M. Waters, (2004) Creating an Institutional Repository: LEADIRS Workbook. Massachusetts: MIT Libraries, pp. 1-134, 2004.
7. Raym Crow, (2002). The Case for Institutional Repositories: A SPARC Position Paper, ARL Bimonthly Report 223 (2002). Available at: http://works.bepress.com/ir_research/7
8. Silva, J.R.d., Ribeiro, C., Lopes, J.C. (2014). Ontology-based multi-domain metadata for research data management using triple stores. In: Proceedings of the 18th International Database Engineering & Applications Symposium
9. http://bazaar.eprints.org
10. http://demoprints.eprints.org
11. http://dspace.nitrkl.ac.in
12. http://dspace.org
13. http://ethesis.nitrkl.ac.in
14. http://wiki.duraspace.org
15. http://wiki.eprints.org
16. http://wikipedia.org
17. http://www.archivematica.org
18. http://www.eprints.org