JOURNAL OF COMPUTING, VOLUME 2, ISSUE 8, AUGUST 2010, ISSN 2151-9617
HTTPS://SITES.GOOGLE.COM/SITE/JOURNALOFCOMPUTING/
WWW.JOURNALOFCOMPUTING.ORG

Institutional Repositories: Features, Architecture, Design and Implementation Technologies

A. O. Adewumi and N. A. Ikhu-Omoregbe

Abstract—Europe is the leading continent in terms of active adoption and use of Digital Libraries – particularly Institutional Repositories (IRs). Africa has not done poorly in this area with a steady increase from 19 repositories in 2008 to 46 in January, 2011 but there is need to raise awareness and channel efforts towards making IRs easily accessible to Africans through ubiquitous channels such as hand-helds and mobile devices. This paper reviews the features, architecture, design and implementation technologies of IRs. In addition, it highlights viable research areas that can be pursued by African researchers in the field of Digital Libraries. It also encourages research efforts to focus on areas that will be beneficial to Africa.

Index Terms—Digital Libraries, DSpace, EPrints, Institutional Repositories

1 Introduction

IN the field of Digital Libraries (DLs) with special emphasis on Institutional Repositories (IRs), Europe is leading in terms of active adoption and use [11]. Research efforts have also been put in place since 2001 through a body named DELOS to investigate future directions of Digital Library research [3].

Africa has not done poorly in terms of adopting and actively using IRs. Statistics in [11] shows that IRs have grown from 19 in 2008 to 46 in January 2011. This accounts for 2% on a worldwide scale. More efforts have to be made however to raise this percentage as it is small compared to Europe's 45%. Also research efforts should be channeled towards making IRs easily accessible to Africans through ubiquitous channels such as hand-held and mobile devices.

An Institutional Repository (IR) is a specialization of a Digital Library (DL). This is inferred from the definitions given by [2], [7] and [8]. From the definitions it can be deduced that IRs and DLs have a common goal of collecting and preserving materials in digital formats.

They also make materials available to a user community. The key difference between the two is that an IR is tailored specifically to capture, preserve and disseminate the intellectual output of a University community or research institution.

According to [8], IRs emerged as a new strategy that allows universities to apply serious systematic leverage to accelerate changes taking place in scholarship and scholarly communication. He further states that many technology trends and development efforts came together to make the strategy possible. Among the factors include: the significant drop in online storage costs, the affordability of repositories; and the establishment of standards like open archives metadata harvesting protocol [8].

Institutional Repositories can be created using IR software. Any institution intending to create an IR must consider the following factors in choosing IR software [1]: software product model (open source software, proprietary software or software service model); features of the software (file formats supported, interoperability-OAI compliance, end-user access to content, API for customizing the software, and persistence of item locator); and technology cost considerations (hardware and servers, operations staff, programming staff, backup and recovery, and preservation).

From the statistics in the Directory of Open Access Repositories [11], there are about 77 known IR software platforms. Of all the platforms, DSpace and EPrints are most popular [9]. This is due to large number of institutions and organizations that employ them in creating institutional/organizational archives. The two platforms are open source and they were built by institutions of higher learning. DSpace was developed by MIT Libraries in collaboration with HP Research Labs while EPrints was developed by the University of Southampton. In this paper, eleven (11) IR platforms were sampled and reviewed in order to highlight the features, architecture, implementation technologies, and design rationale of IRs. The sampled platforms are as follows: CONTENTdm, Digital Commons, DigiTool, DSpace, EPrints, EQUELLA Repository, Greenstone, Islandora Fedora, intraLibrary, Open Repository and Zentity. In addition, the paper identifies current and future trends in IR platform development.

- *A.O. Adewumi is with the Department of Computer and Information Sciences, Covenant University, Ogun State, Nigeria. E-mail: adewunmi@covenantuniversity.com.*
- *N.I.Omoregbe is with the Department of Computer and Information Sciences, Covenant University, Ogun State, Nigeria. E-mail: nomoregbe@gmail.com.*

2 Features of IRs

2.1 Open Source or Proprietary

An IR platform can either be open source or proprietary. When it is open source, it can be downloaded and installed out of the box free of cost. The codes that make up the platform are also publicly available and institutions intending to use such a platform can customize the platform to suit their purpose. On the other hand, when it is proprietary, the proprietor has the sole right to the platform and its codes and will only install and administer for institutions at a cost. Six of the platforms we sampled are free and open source namely: DSpace, EPrints, Greenstone, Islandora Fedora, intraLibrary and Zentity. The other five are proprietary. One point to note when choosing an IR platform (especially one that is open source) is that though it is free and open source, there might be some hidden costs especially when it comes to carrying out customizations.

2.2 Software or Hosted Service

Apart from being open source and proprietary, IR platforms can come as either software or as a hosted service. As software, they can be downloaded and installed either free of charge or at a cost depending on the platform chosen. However, as a hosted service, the institution or organisation (client) intending to use such a platform will subscribe to the proprietor who acts as a service provider. The client will give specification of features they desire to be present in their IR web pages while the service provider will then compile these specifications and build the IR to the client's taste. This is done at a cost. The service provider will also be responsible for administering the IR so that the client can focus on populating the IR. Of all the platforms we reviewed only two were a hosted service namely: Digital Commons and Open Repository. The rest of them come as software.

2.3 Support

This refers to help that is provided to users of a particularr IR platform. It is of varying kinds. Support can be provided through a community. This is found in some of the open source platforms we sampled namely: DSpace, EPrints and Islandora Fedora. In this kind of support, users of the particular IR platform join the platform's community mailing list and can share problems they encounter while using the IR platform. Other members who might have encountered and overcome similar challenges then help in troubleshooting and tackling the problem. The community members also help to update the IR platform with new features and functionalities on a regular basis. It was discovered during the course of this review that IR platforms with community support do not charge for platform updates. Another kind of support is the direct support. Here the user of an IR platform can get help directly from the proprietors of the platform. This kind of support is present in Open Repository and Zentity. Some other IR platforms offer support as a service that is paid for. They include: CONTENTdm, Digital Commons, DigiTool, DSpace, EPrints, Islandora Fedora and intraLibrary.

2.4 Content

The IR platforms we sampled can store items of various formats including audio files, video files and images.

2.5 Metadata Formats

Metadata are records that refer to digital resources available across a network [6]. Metadata in the context of IRs can be referred to as data that helps to describe the digital resources (content) stored in IRs. Some standard metadata formats that are supported on IR platforms include: Dublin Core (DC), Qualified Dublin Core (QDC), METS and MARC. In our sample IR platforms, DC is supported on all platforms. QDC is supported on all platforms but EPrints and Zentity. METS is also supported on all platforms but Digital Commons, intraLibrary and Open Repository. MARC is only supported on DigiTool, EQUELLA Repository and Islandora Fedora.

2.6 User Interface Functions

In the sampled IR platforms, two key user interface functions were identified as being common to all namely an End-user Deposition Interface and a Multi-language support interface. An End-user Deposition Interface is one that allows an end-user (e.g. faculty at a university) to deposit items (e.g. preprint papers) in an IR. The Multi-language support function allows an IR support more than one language especially when the expected audience of an IR is non-English speaking.

2.7 Advanced Searching

IRs depending on the purpose of their use can sometimes contain a large number of records (up to a million records). As a result, most IR platforms (particularly those sampled in this work) come with a search facility. The search can be both simple and advanced. A simple search is field specific while an advanced search can include Boolean logic and sorting options.

2.8 Default Subject Classes

This refers to how items in an IR are classified. It is closely related to how books are catalogued in a library. From the sample IR platforms we examined, it was discovered that very few of the platforms namely EPrints and intraLibrary have Default Subject Classes. This means that most IR platform developers leave the classification of IR items to the repository administrator. EPrints and intraLibrary have Library of Congress Classification as their default subject class. In addition, intralibrary also has Dewey Decimal Classification (DDC) as default.

2.9 Syndication

According to [12] it is the controlled placement of the same content on multiple partnering sites. There are two types of syndicated content namely [12]: RSS or Atom feeds and Full Content. Some of the IRs sampled support either of RSS or Atom feeds. In some instances they support both. RSS is found in all the platforms we sampled except CONTENTdm and DigiTool while Atom is found in DSpace, EPrints, EQUELLA Repository, Islandora Fedora and Zentity.

2.10 User Validation

Depending on the type of restriction set on items in an IR, just about any person can download and view IR content especially in IRs where Open Access is supported. However, for a user to deposit items in an IR s/he needs to be registered on that IR. This can be done by filling an electronic form that will among other things request for a preferred user name and password. This feature is supported by all the platforms we examined. IRs also allow for other means of authentication such as LDAP, Shibboleth and Athens. LDAP Authentication is supported on all the sampled IR platforms. However, Shibboleth and Athens are supported on some of the platforms.

2.11 Web 2.0

This is a term used to describe the Web as we have it today. According to [4] it has evolved from being just an information source to becoming a participatory Web where users can actively engage in generating content. As an information source (Web 1.0), the Web consisted of text, images and hyperlinks. The Web as we know it now has evolved to include: wikis, blogs, bookmarking tools and the likes. With Web 2.0 come concepts like: tagging, comments, ratings, reviews, bookmarks and share this functionality on websites. IR platforms are gradually adopting these concepts and implementing them. From the sampled IR platforms, DigiTool, EQUELLA Repository and Islandora Fedora have fully implemented these features. The other IR platforms have one more of the features implemented. CONTENTdm however, has none of these features. The proprietors have it in mind though to incorporate these features in subsequent versions of the software.

2.12 Statistical Reporting

Items are placed in IRs for visibility and access to a wide range of audience. As such faculty who deposit items would want to know how frequently their deposited items are downloaded. It would also be of interest to some repository managers or even first time visitors of an IR web page to know the exact number of items in an IR. As a result of this, all the IR platforms we sampled have Top Downloads functionality as well as a Count functionality that enables one to know the number of items in an IR archive.

2.13 Machine-to-Machine Interoperability

This has to do with the level to which various IR platforms are able to interact and share information. In other to achieve interoperability, certain standards must be adhered to among which is OAI-PMH, OAI-ORE, SWORD, SWAP, RDF, RoMEO Integration, OAI-PMH Harvesting. Among the IR platforms sampled, EQUELLA Repository and Islandora Fedora fully support these standards. In addition, it was discovered that all the platforms in particular support OAI-PMH.

2.14 Administrator Functions

All the IR platforms sampled allow an IR administrator to carry bulk imports, bulk exports and also customize IR workflows. Bulk imports have to do with bringing content en-masse into an IR from an external source. The reverse is the case for bulk exports.

3 ARCHITECTURE OF IRS

A close examination of the sampled IR platforms reveals that the architecture of IR platforms can be classified into two namely: Open and Closed architecture.

3.1 Open Architecture

It is one that is modular, extensible and can be accessed and modified by members of the public. An open architecture can be contributed to by a group of persons not necessarily the platform developers. Open source IR platforms usually possess this type of architecture. The open architecture of the sampled IR platforms can be further sub-divided into three-tier architecture and Plug-in architecture. Most IR platforms possess the three-tier architecture [10] except for EPrints that has a flexible plug-in architecture for developing extensions [5]. The next two paragraphs discuss the architecture of the two most popular IR platforms sampled namely: DSpace and EPrints.

The DSpace architecture is a straightforward three-layer architecture, including storage, business and application layers, each with a documented API to allow for future customisation and enhancement [10].

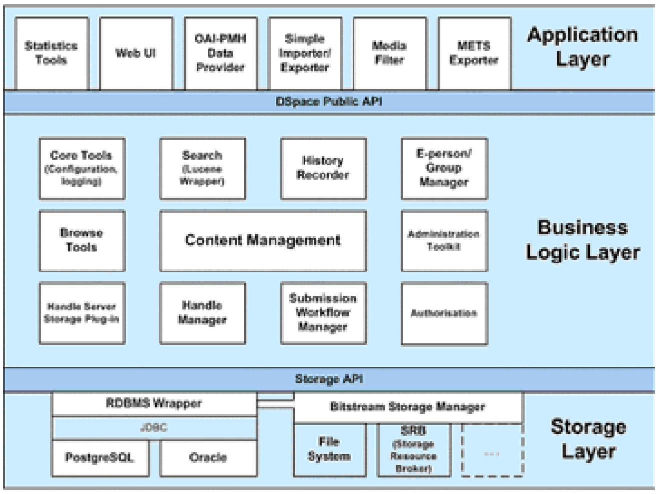

Fig 1: DSpace 3-tier architecture

EPrints provides a flexible plugin architecture for developing extensions [5]. It is in fact a generic plugin framework with a set of plugins that implement the functions of a repository. Most of the dynamic Web pages in EPrints are actually screen plugins. Also, all import/export options are implemented as plugins. In addition, all input components in deposit workflow are plugins. It gives plugin developers many examples to work from. A diagram to depict the EPrints architecture is shown in Figure 2.

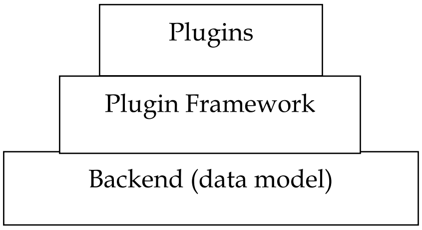

Fig 2: Architectural Framework of EPrints [5]

3.2 Closed Architecture

It is one that is not accessible to the public as a result it cannot be extended or modified by anyone but the proprietor. Proprietary IR platforms such as Digital Commons and Open Repository have this type of architecture.

4 Design Rationale of IRs

A close examination of the sampled IR platforms further reveals that the rationale behind the development of IR platforms is to ensure the following:

4.1 Flexibility

The sampled platforms were all able to store items of various formats including audio, video and images files. This is very vital especially as IRs platforms are being used for purposes other than what they were originally meant forstoring the research output of a university community.

4.2 Accessibility

Items stored in an IR should be accessible through various scholarly search engines such as Google Scholar and Scirus. IR platform developers put this into consideration when developing their IR platforms. For instance, an IR that is deployed using EPrints can be made accessible by registering it on Google Scholar, Scirus, Registry of Open Access Repositories (ROAR) and OpenDOAR.

4.3 Interoperability

IRs built on dissimilar platforms should be able to interact and share information. This is vital especially in promoting Open Access.

4.4 Standards-Based

The sampled IR platforms were developed using widely accepted standards such as OAI-PMH. This promotes interoperability.

4.5 Security Options

The platforms sampled provide institutions with the option of determining who has access to what content on their IR web pages.

5 Implementation Technologies of IRs

Based on the sampled IR platforms, the implementation technologies have been classified into three namely: scripting language, database and operating system.

5.1 Scripting Language

Java, Perl, PHP, JavaScript and AJAX are some of the major scripting languages used in developing the IR platforms sampled. Others include Extensible Stylesheet Language Transformations (XSLT) and .NET. Some of the platforms are written entirely in one scripting language while others are written using a combination of scripting languages. CONTENTdm and Digital Commons for instance are written entirely in PHP and Perl respectively. DigiTool, DSpace, EQUELLA Repository, Greenstone, Islandora Fedora and intraLibrary are written in Java but combine some of the aforementioned scripting languages. Also, EPrints is written in Perl but combines with JavaScript, AJAX and XSLT. Zentity is the only platform written using .NET and the reason is not far-fetched. It was developed at Microsoft.

5.2 Database

MySQL, Oracle, PostgreSQL and Microsoft SQL Server are the major database systems used by the sampled IR platforms. Some of the platforms are compatible with only one of the database systems while some others are compatible with two or more of the database systems. MySQL is compatible with EPrints, Islandora Fedora and intraLibrary. Oracle is compatible with DigiTool, DSpace, EPrints, EQUELLA Repository and Islandora Fedora. PostgreSQL is compatible with Digital Commons, DSpace, EPrints, EQUELLA Repository and Islandora Fedora. Microsoft SQL Server is compatible with EQUELLA Repository, Islandora Fedora and Zentity. More recently cloud storage has been introduced in EPrints and Islandora Fedora.

5.3 Operating System

Linux, UNIX, SOLARIS, Windows and Mac OS X are the operating systems on which the sampled IR platforms run. Some of the IR platforms run on only one. For example, Digital Commons runs on Linux and Zentity runs on Windows. The other platforms run on two or more of the operating systems. It was noticed that most of the IR platforms that run on all the aforementioned operating systems were written in Java and so are platform independent. An exception to this is EPrints. EPrints runs on all the operating system platforms and yet is written in Perl.

6 Possible Research Directions

6.1 IR Architectures

As new features and functionalities emerge in IRs, the present 3-tier architecture that most IR platforms possess may become inadequate. Although the plugin architecture of EPrints is a welcome development, there is need to explore more novel architectures, particularly the Grid and peer-to-peer approaches and several forms of service architecture [13].

6.2 Mobile Access

Of all the eleven (11) IR platforms reviewed, only Greenstone supports access via mobile devices. This is obviously an opportunity that can be explored by budding African researchers. Another motivation for this is that mobile phones have really penetrated the African landscape (particularly sub-Saharan Africa) at an increased rate over the past decade [14] giving rise to new possibilities as discussed in [15]. Therefore, making IRs accessible on a mobile phone will among other things help to create awareness of its existence and potential among Africans.

7 Conclusion

This paper has discussed the features, architecture, implementation technologies and design rationale of IRs and also highlighted possible research opportunities in the field. It is believed that it will help enlighten persons (particularly African researchers) intending to do research in this field.

Acknowledgment

This work was supported in part by a grant from Covenant University Centre for Research and Development (CUCERD).

References

[1] M.R. Barton and M. M. Waters, *Creating an Institutional Repository: LEADIRS Workbook*. Massachusetts: MIT Libraries, pp. 1-134, 2004.

[2] Raym Crow, “The Case for Institutional Repositories: A SPARC Position Paper” *ARL Bimonthly Report 223* (2002). Available at: http://works.bepress.com/ir_research/7

[3] ***, “Digital Libraries: Future Directions for a European Research Programme,” *DELOS Brainstorming Report*, San Cassiano, Italy, June 2001 http://delos-noe.iei.pi.cnr.it/activities/researchforum/Brainstorming/brainstorming-report.pdf

[4] B. Decrem, “Introducing Flock Beta 1," *Flock Official Blog*. http://www.flock.com/node/4500 2006

[5] ***, "Advanced Customization: Scripting EPrints," *EPrints Training Course.* http://www.eprints.org/software/training/programming/api_techniques.pdf 2010

[6] R. Heery, “Review of Metadata Formats”, *Program*, vol. 30, no. 4 pp. 345-373, 1996

[7] L. Candela, D. Castelli, N. Ferro, Y. Ioannidis, G. Koutrika, C. Meghini, P. Pagano, S. Ross, D. Soergel, M. Agosti, M. Dobreva, V. Katifori and H. Schuldt “The DELOS Digital Library Reference Model – Foundations for Digital Libraries”, Version 0.98 http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf 2007

[8] C.A. Lynch, “Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age,” *ARL*, no. 226 February 2003: 1-7 http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

[9] A. Sale, “The Key Things to Know,” *University of Tasmania EPrints Repository*, http://eprints.utas.edu.au/223/01/NZ_Workshop_PDF.pdf 2005

[10] M. Smith, M. Barton, M. Bass, M. Branschofsky, G. McClellan, D. Stuve, R. Tansley, and J. H. Walker, “An Open Source Dynamic Digital Library”. *D-Lib Magazine*, vol. 9, no. 1, January 2003.

[11] ***, “Directory of Open Access Repositories” http://www.opendoar.org 2011

[12] ***, “White Paper on Internet Content Syndication,” http://www.internetcontentsyndication.org/downloads/whitepapers/content_creation.pdf

[13] ***, “Future Research Directions,” 3rd DELOS Brainstorming Workshop Report, Corvara, Italy, July 2004 http://www.delos.info/files/pdf/events/2004_Jul_8_10/D8.pdf

[14] J. C. Aker and I. M. Mbiti, “Mobile Phones and Economic Development in Africa”. *Journal of Economic Perspectives*, vol. 24, no. 3, pp. 207-232 2010 doi=10.1257/jep.24.3.207

[15] J. Hellstrom, “The Innovative Use of Mobile Applications in East Africa”. *Swedish International Development Cooperation Agency* http://upgraid.files.wordpress.com/2010/06/sr2010-12_sida_hellstrom.pdf

A. O. Adewumi obtained a B.Sc in Computer Science from Covenant University in 2008 and currently works as a Graduate Assistant at the Department of Computer and Information Sciences of Covenant University while also studying for an M.Sc in Computer Science at the same institution. His research interest is Digital Libraries with special emphasis on Institutional Repositories. He is a Sun Certified Java Programmer.

N. A. Ikhu-Omoregbe holds a B.Sc degree in Computer Science from the University of Benin, Benin City, an M.Sc. degree in Computer Science from the University of Lagos, and a PhD degree in Computer Science from Covenant University, Ota, Nigeria. His research interests include: Software Engineering, Mobile Computing, Multimedia technologies, Mobile Healthcare and Telemedicine Systems, and Soft Computing. He currently lectures in the Department of Computer and Information Systems, Covenant University, Ota, and has taught at Baden-Wurttemberg Cooperative State University, Heidenheim as a visiting lecturer in the area of e-Health Systems.