Open Source ILS (IR)
An over view on Dspace

Speaker: Sher Afzal Khan Pakistan

Introduction

In the digital age, the amount of data produced is growing exponentially. Governments and institutions can no longer rely on old methods for storing data and passing on the knowledge to future generations. Digital data preservation is a mandatory issue that needs proper strategies and tools. With this awareness, efforts are being made to create and perfect software solutions capable of responding to the challenge of properly preserving digital information.

Introduction Cont….

Information preservation can simply be defined as the set of processes to store, index and access information

In recent years, the creation of digital content has grown exponentially. Gantz and Reinsel report that the so called digital universe will grow from 2005 to 2020 by a factor of 300, from 130 exabytes to 40,000 exabytes.

Smartphones, with all their data sensors, namely photo and video recording capabilities, are also major contributors to the current massive production of data. The Internet of Things (IoT) is poised to generate increasing amount of data, even if IoT middleware can help by reducing the volume of data to store and preserve

In fact, major trends like Big Data have fostered the perception of digital data as valuable assets, strengthening the need for digital data preservation and henceforth for proper digital repositories

Free and Open-Source Software

(FOSS) is software that can be classified as both free software and open-source software.

That is, anyone is freely licensed to use, copy, study, and change the software in any way, and the source code is openly shared so that people are encouraged to voluntarily improve the design of the software. This is in contrast to proprietary software, where the software is under restrictive copyright and the source code is usually hidden from the users.

Richard Matthew Stallman (born March 16, 1953), often known by his initials, rms, is an American free software movement activist and programmer. He campaigns for software to be distributed in a manner such that its users receive the freedoms to:

1.Use 2.Study 3.Distribute 4.Modify

Software that ensures these freedoms is termed free & open Source software.

Libraries and OSS

- OSS “model is culturally quite similar to the values embodied in libraries”
- Libraries and open source software are a natural fit.
- “OSS is solving problems for libraries of all types and is helping to get work done for people of all technical abilities”
- OSS offers cheap alternatives to expensive commercialized solutions considered, Free and open Source Software a useful and significant tool that can be helpful to solve many problems faced by the people of developing countries.
- The report identified low cost, anti-piracy campaign, security and technological independence as factors behind the expansion of FLOSS in the developing world.
- “by standing up for open source within our libraries we stand up for libraries as evolving organizations ready for the challenges of the future”

Open Source Integrated & IR Library Systems

Koha
MicroLCS
Emilda
Evergreen
FireFly
GNUTeca

Avanti
OpenBiblio
PhpMyLibrary
PMB
PYTHEAS
WEBLIS

IR Software

Dspace
Archimede
DAITSS
Fedora
Xena

Green Stone
Archivematica
E Prints
RODA
Invenio

DSpace

DSpace is an open source software application developed and supported by a vibrant user community.

DSpace is a Digital Repository Software, created as a joint project of MIT Libraries and the Hewlett-Packard Company, and publicly released in November 2002 as Open-Source Software.

DSpace is one of the open source software platform to store, manage and distribute the collections in digital format. As much of the world's content is now being developed and disseminated in digital format, the DSpace software supports next-generation digital archiving that is more permanent and shareable than current analog archives.

DSpace Cont...

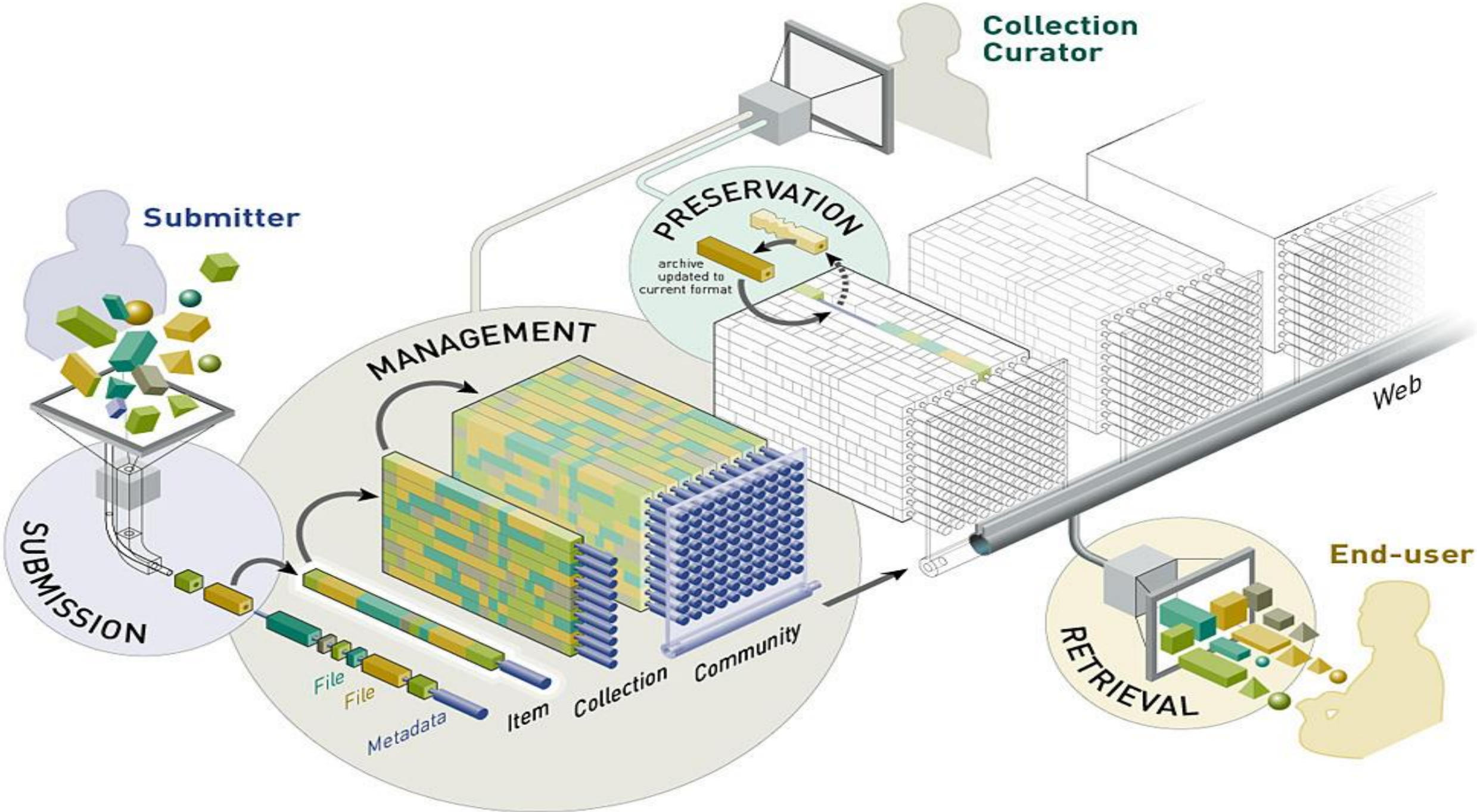
With 1,200+ known installations in over 100 countries, the DSpace out-of-the-box application makes it simple to launch an open repository to deliver digital content to end users.

DSpace includes features and tools to manage, preserve and provide access to content on the Web. DSpace can store any type of digital content, offers built-in workflows for content submission and review, and can be customized to fit your needs.

DSpace Cont...

The most common use of the DSpace software is by academic and research libraries as an open access repository to manage and provide access to faculty and student output.

Other types of organizations using DSpace, include national libraries, museums, cultural institutions, governments and businesses that host and manage everything from special collections/subjects, research, datasets to audio/video materials.


Collection
Curator
Submitter
PRESERVATION
archive
updated to
current format
MANAGEMENT
Web
SUBMISSION
File
File
Metadata
Item
Collection
Community
End-user
RETRIEVAL

DSpace Functions

Captures

- Digital research material (any format)
- Directly from creators (e.g. faculty)
- Large-scale, stable, managed long-term storage

Describes

- Descriptive, technical, rights metadata
- Persistent identifiers

Distributes

- Via WWW, with necessary access control

Preserves

Possible Content

- Preprints, articles
- Technical Reports
- Working Papers
- Conference Papers
- E-theses
- Images
  - visual, scientific, etc.
- Audio files
- Video files
- Learning Objects
- Reformatted digital library collections

OAI Support

- The Open Archives Initiative has developed a Protocol for Metadata Harvesting (OAI-PMH) . This allows sites to programmatically retrieve or 'harvest' the metadata from several sources, and offer services using that metadata, such as indexing or linking services. Such a service could allow users to access information from a large number of sites that are collated in a central catalog.
- DSpace exposes the Dublin Core metadata for items that are publicly (anonymously) accessible. Additionally, the community and collection structure is also exposed via OAI-PMH's 'sets' mechanism. OCLC's open source OAICat framework is used to provide this functionality

DSpace's OAI-PMH service exposes deletion information for withdrawn items

DESIGN

Information Model

Communities
- Research units of the organization

Collections (in communities)
- Distinct groupings of like items

Items (in collections)
- Logical content objects
- Receive persistent identifier

Bitstreams (in items)
- Individual files
- Receive preservation treatment

Information Model

Versioning

- Item “versions” can be
  - All instances of a work in different formats
    - E.g. the XML, PDF, and PostScript versions
  - All editions of a work over time
    - Official changes (e.g. addenda or new release)
    - Periodic snapshots (e.g. web sites)
- Metadata lists all available versions of items

Communities

Research units of the organization

- Schools, Departments, Research Labs, Research Centers, Programs, etc.
- Individuals

Community “home page” with logo, custom description, etc.

- Or contract with library

Communities

Local, distributed policy decisions

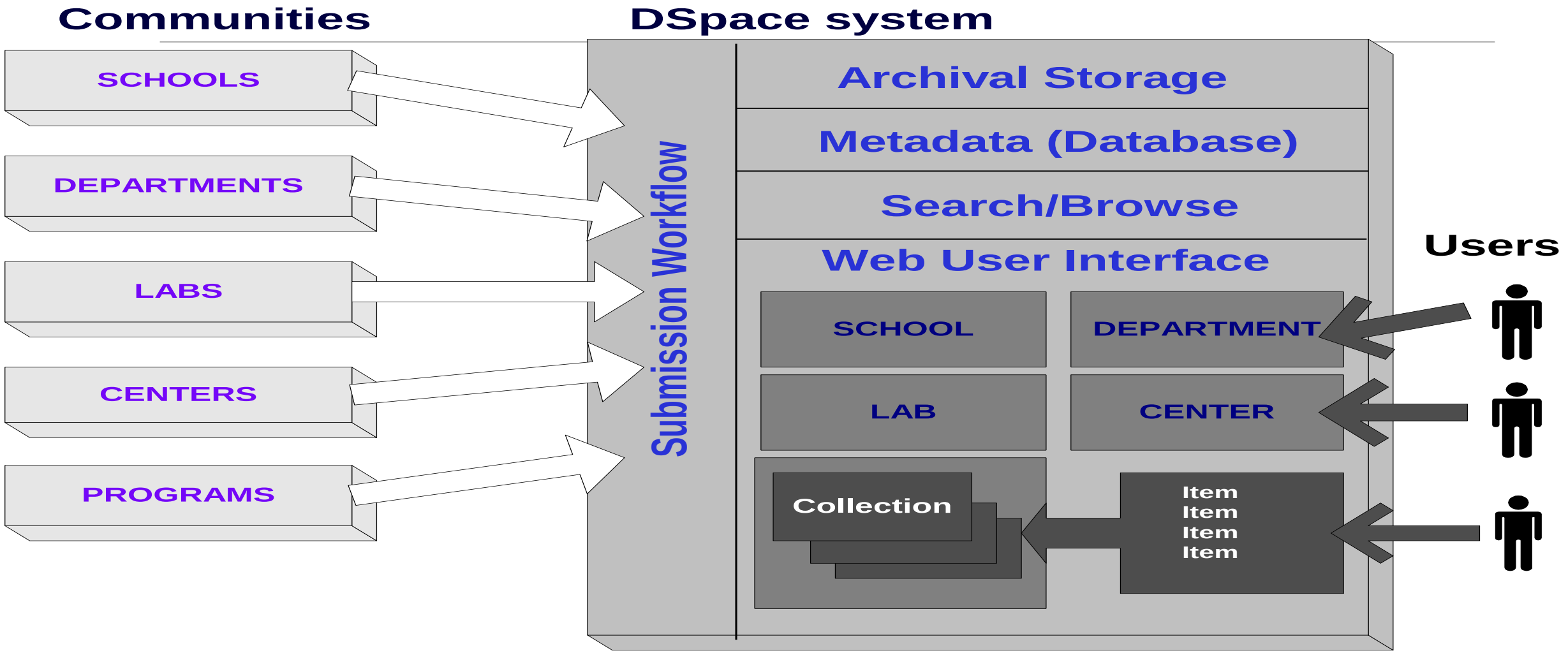
- Who can contribute, access material
- Submission workflow
  - Submitters, approvers, reviewers, editors
- Collections definition, management

Local, distributed production work

- Communities supply metadata, files

Partnership between library and communities

Communities



Metadata

DSpace uses a qualified Dublin Core metadata standard for describing items intellectually (specifically, the Libraries Working Group Application Profile).

Only three fields are required: title, language, and submission date, all other fields are optional. There are additional fields for document abstracts, keywords, and technical metadata and rights metadata, among others.

This metadata is displayed in the item record in DSpace, and is indexed for browsing and searching the system (within a collection, across collections, or across Communities).

User Interface

DSpace's current user interface is web-based. There are several interfaces:

- one for submitters and others involved in the submission process,
- one for end-users looking for information, and one for system administrators.

Technology Platform

DSpace was developed to be open source, and in such a way that institutions and organizations with minimal resources could run it. The system is designed to run on the UNIX platform, and comprises other open source middleware and tools, and programs written by the DSpace team.

All original code is in the Java programming language. Other pieces of the technology stack include a relational database management system (PostgreSQL), a Web server and Java servlet engine (Apache and Tomcat, both from the Apache Foundation), Jena (an RDF toolkit from HP Labs), OAICat from OCLC, and several other useful libraries.

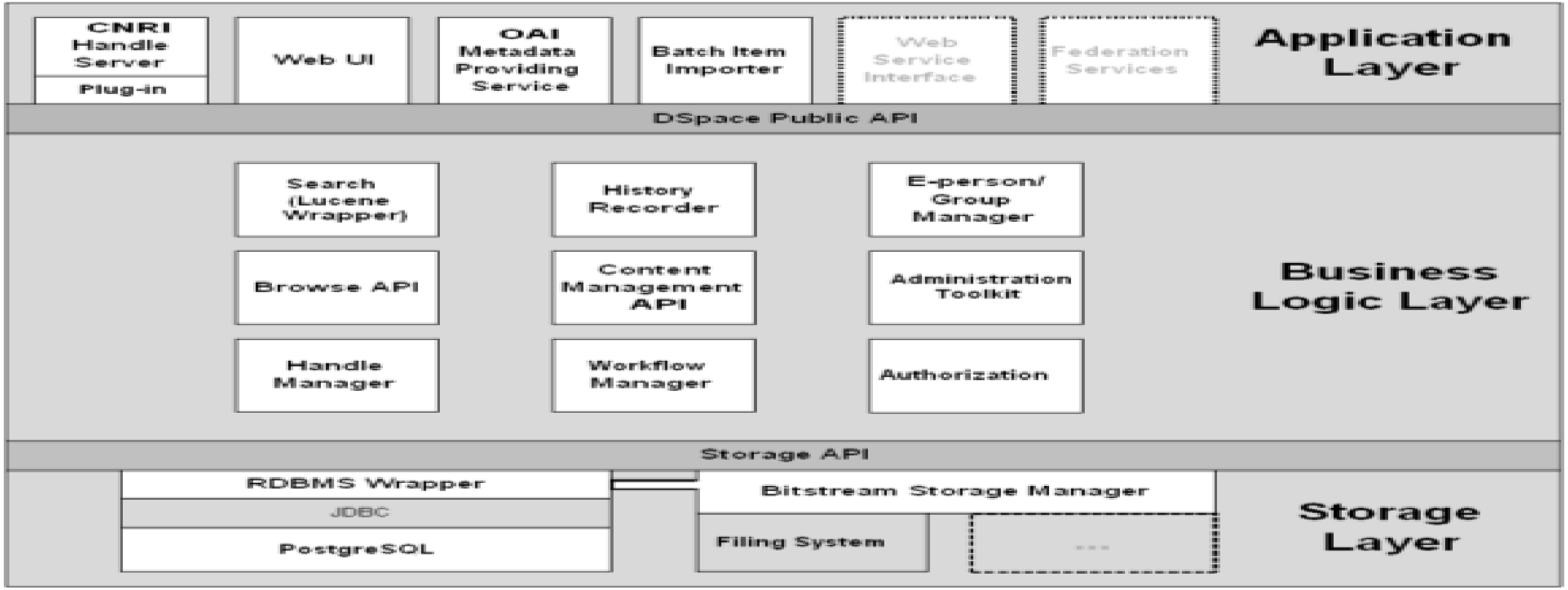
System Architecture

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

The storage layer is implemented using the file system, as managed by PostgreSQL database tables.

The business layer is where the DSpace-specific functionality resides, including the workflow, content management, administration, and search and browse modules. Each module has an API to allow DSpace adopters to replace or enhance that function as desired.

System Architecture

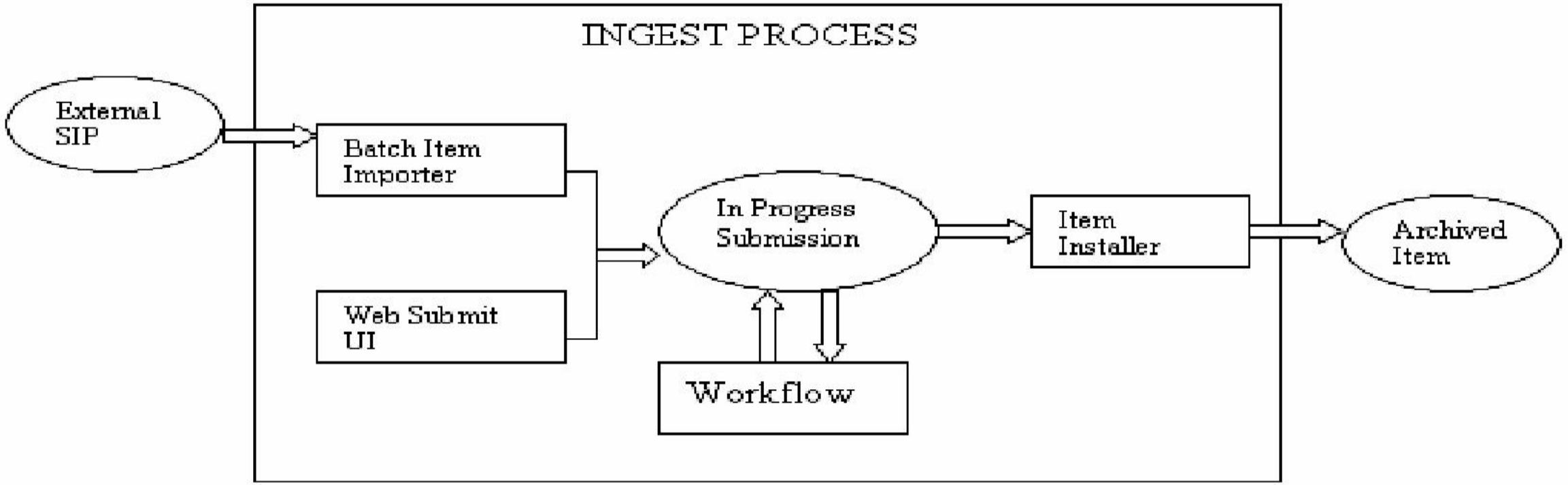


DSpace Ingestion

DSpace is the software that serves as a repository and stores digital content. In a system with such a goal, perhaps the most critical aspect of the system is how that data enters the system.

This occurs mainly in two ways in DSpace. The web based UI for the software allows a user to submit items to collections as long as they are logged in as a registered user. When users log in, they go through a configurable workflow where they upload and describe their submissions

DSpace Ingestion Preocess



DSpace Workflow

The DSpace submission workflow system is a critical part of the DSpace architecture that allows submission, processing, and final addition of content to the live repository.

DSpace's underlying model includes E-People, users who have registered with the system and have certain authorizations, roles, rights, and privileges that translate abilities to complete certain tasks within the Dspace system.

A typical submission begins with the system asking the user a couple of questions about digital document to be added in the repository and number of files involved in the submission.

DSpace Workflow

Workflow Step	Description
Describe	**User enters metadata about the document (s) they are submitting, including but not limited to author, title, keywords, and a description**
Upload	**The user selects and uploads the files on their local machine that they like to upload as part of the submission. Each file's type is identified by the system and the user verifies the file type**
Verify	**An overview of all details of the submission is given including a summary of the entered metadata and the files involved in the submission.**
License	**The user is shown and must agree to the license the system administrator has assigned to submit content for this collection**
Complete	**The user's actions in the submission process are complete. Based on the workflow steps set for the collection, the item may immediately be added to the collection or have to be reviewed by system administrators before its addition to the collection.**

Communities and Collection Process

Items can be placed in a collection

Example:

- A collection of theses
- A collection of reports
- A collection of articles
- A collection of e-mails

Collections are part of a community or sub-community

Communities can be divided into sub-communities, which can be further sub-divided

Examples of communites

- Social Sciences
- Natural Sciences
- Humanities
  - History
  - Psychology
  - Philosophy
    - Logic
    - Metaphysics

Collections in Dspace

Under each community or sub-community, you should have collection(s).

It is the collections that contain items (digital documents)

Communities or sub-communities can not have items directly under them.

Collections can be organized by type of documents

For example,

- Theses
- Articles
- Photographs
- Presentations etc.

Again, it depends on users expectations

WHY DSAPCE

Why Dspace…

Largest Community Of Users And Developers Worldwide

- Ability To Choose the Default Language
- Used By Educational, Government, Private and Commercial Institutions
- Can be Installed out of the box

Completely Customizable To Fit One's Needs

- User interface
- Ability To Customize the Metadata
- Ability to configure Browse and Search
- Configurable database
- Ability To Choose the Default Language

Why Dspace…

- Online access to your digital assets
- Full-text search
- Optimized for Google Indexing
- OpenURL Support
- Support for modern browsers
- OAI Support

Who Uses DSpace?

International Users

MIT Libraries
- https://dspace.mit.edu/
- University of Hong Kong
  - https://hub.hku.hk/handle/10722/143269\
- Texas Digital Library
  - https://www.tdl.org/repositories/
- The World Bank
  - https://openknowledge.worldbank.org/
- Harvard University
  - https://dash.harvard.edu/contact
- Cambridge University
  - http://www.lib.cam.ac.uk/repository/

Conclusion

One of the leading uses for DSpace is as an institutional repository. DSpace followed the librarian's inclination to create a system that would be as easy as possible to implement and use, rather than push strictly in the direction of digital library research from which a more flexible system might have emerged.

With all the important functions, supporting all kind of documents, customization according to the organization needs and large number of community Dspace is the best example for any type of organization.

OSS Contributions –Pakistan

- Organized 130 + Workshop so far on OSS
- Custom Live DVD of Koha and Dspace (LIBSOL) With latest features (4 versions)
- Facebook Group " Pakistan Koha and Dspace LIS Community
- Blogs: Blog.fullstack.com.pk
- Pioneer Country Ambassador For Dspace
- Worked as members for the establishment of Pakistan Research Repository using Dspace
- Founder and President Pakistan Koha Community –Koha Pakistan
- Organized 02 International Conferences
- Patriated 4 time as Speaker in International Open Source Conference
- Winner of Best Library award, Emerald group of publication and IEEE for Pakistan
- Implanted open Source Software in 55 + organization i.e. Universities, Public and Private organization, National Assembly, POF and many other prestigious organization.
- Technical Committee member of PASTIC Union Catalogue Team

References

- http://www.dspace.org
- http://www.sourceforge.net/projects/dspace
- http://wiki.dspace.org
- http://nsdl.niscair.res.in (National Science Digital Library)
-]OCLC Research OAICat Open Source Project. http://www.oclc.org/research/software/oai/cat.shtm
- Open Archives Initiative. http://www.openarchives.org/
- www.kohapaksitan.org
- https://blog.fullstack.com.pk/
- Rosa, Carlos André, Olga Craveiro, and Patricio Domingues. "Open source software for digital preservation repositories: a survey." *arXiv preprint arXiv:1707.06336* (2017).

THANKYOU FOR YOUR ATTENTION
ANY QUESTIONS?

Sher Afzal Khan

President & Founder Koha Pakistan, Vice Chairman Read Pakistan, Lead Trainer of Koha & DSpace, Entrepreneur and Lead Consultant

www.kohapaksitn.org | www.fullstack.com.pk |kohapaksitan@gmail.com