Indian Virtual Herbarium Project: Implementing an institutional knowledge repository as a digital archive – Design, development, solution architecture and implementation

Ramesh Singh

is a senior scientist at NIC, has worked extensively in the OS, Database Management Systems and other system software products organizations. He has studied at IIT Kharagpur and at IIT Delhi. He holds an MTech from IIT Delhi. His present research interests are in mobile, pervasive computing and human computer interfaces.

Kush Sharma

holds a BTech (CSE) from NIT Hamirpur and an MBA (Systems). He is currently working for NIC in the capacity of Java Architect. He is an experienced hand in JAVA programming and has a flare for Technology research areas, particularly JAVA architecture, digital preservation and metadata research.

ABSTRACT The Indian Virtual Herbarium (IVH) portal is proposed as a customization of the *DSpace Digital Library System* for archiving. The portal will consist of a user-friendly interface. The classification hierarchy will consist of the Herbariums at the top level, collection plants and then the actual plant image stored in the tiff/jpeg format. The images are assumed to be scanned in an appropriate way with the desired resolution. This paper describes the solution architecture based on the digital library product DSpace, and also outlines the requirements of System Software. The three major missions of this paper are to provide technology services for the IVH, DH and Production Projects; to deal with the standards of information organization; and to build the National Botanical Digital Archives Resource Center. The set up of the collaboration and communication mechanisms involve an integrated searching system; a knowledge management system; and a digital preservation mechanism. Backup services also have to be provided

Journal of Digital Asset Management (2009) 5, 55–74. doi:10.1057/dam.2008.56

Keywords: Herbarium; digitization; metadata; preservation; archives; DSpace

INTRODUCTION

The system will provide a means to browse the Herbarium and perform a search based on certain criteria such as plant details, the taxonomy, contributors and so on. The system will also provide an ability to add the images to the collection. Different users can be assigned

the role of adding, viewing or modifying the images, which in DSpace jargon are called 'items.' There will be users who can simply browse or search for images and view them. The image will be rendered on the basis of the format of the image; the supported formats will be jpeg, gif, tiff, and so on. Once the image has

Correspondence:

Ramesh Singh
National Informatics
Centre, A-Block, CGO
Complex, Lodi Road, New
Delhi 110 003, India

been captured, it will show under the particular Herbarium that it has been ingested. The metadata associated with the image will be stored in the PostgreSQL.¹ The database and the images will themselves be stored in a flat file system.

Although still in its initial phase, this project will create a case of innovative use of the groundbreaking Open Source Digital Library product known as DSpace, a joint venture of MIT and HP Labs,² to develop a content-rich digital repository by connecting researchers fully to the ICT designers and librarians/curators.

The digitalization of these collections involves a wide variety of content and data types, reflective of the many research specialties. Therefore, intensive information technology and resource-sharing information systems will be used to bring coherence to the digital preservation service.

With the constructions of digital contents and the database, the digital archives initiatives not only allow university faculty and students drawing upon the collections to enrich learning and researches, but also serve a wider audience through Internet access. Furthermore, the digital preservation of archives will be engaged in advancing knowledge innovation, facilitating value-added creation and sustaining national competitiveness.

IVH: AS AN INSTITUTIONAL REPOSITORY

IVH will function as a Government of India institutional repository – a set of services that a government offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members.

Elements of digital archives collections

- Botanical Collections of Herbarium Type
 - Type Specimen (# Nos.);
 - Documents of Type Specimen (#Nos.);
 - General Specimen (#Nos.);
 - Flora of India (Volume Details and so on).
- Specimen Metadata
 - plant details (phenology, precode, typequalifier);
 - taxonomy (family, genus, species, intraspecific name, intraspecific rank);

- coverage (country, district, region, maxaltitude, latitude, longitude);
- contributors (collector, determiner, donor);
- date (accessioned, collected, donated, determined).
- Submission Workflow
 - five metadata pages to describe the plant specimen;
 - each page caters to a different category given in Specimen metadata;
 - upload step, where file is uploaded;
 - verify step, where metadata is verified.
- Browse
 - by accession number;
 - by collector;
 - by family;
 - by date.

Search

- Advanced Search by
 - plant details;
 - taxonomy;
 - coverage;
 - contributors;
 - date.
- Simple Search

Searches all of DSpace for the nearest match and even provides full text search. This content digital archive system will store individual data structure, metadata standards, management policy and search interface.

INTRODUCING 'DSpace'

DSpace² is a groundbreaking digital repository system that captures, stores, indexes, preserves and redistributes an organization's research data.

What can DSpace do?

Jointly developed by Massachusetts Institute of Technology Libraries and Hewlett-Packard Labs, the DSpace software² platform serves a variety of digital archiving needs.

Research institutions worldwide use DSpace to meet a variety of digital archiving needs:

- Institutional Repositories (IRs);
- Learning Object Repositories (LORs);
- eTheses;
- Electronic Records Management (ERM);
- Digital Preservation;

- Publishing;
- and more.

Elements of the digital repository

- Design
 - broaden public access;
 - personalize use.
- Services
 - searching;
 - storage;
 - usage;
 - documentation.

Digital library product used

DSpace v.1.4.1

- Open Source
 - Java;
 - related database/SQL.
- APIs;
- ‘handles’ as identifiers;
- Open Archival Institute – Preservation Metadata Harvesting (OAI-PMH);
- management.

DSpace data model

- Communities
- Collections
- Items
- Bit-streams

DSpace-BASED INSTITUTIONAL REPOSITORY IMPLEMENTATION

- System Architecture
- Data Ingestion
- Metadata
 - Harvesting
 - Mapping
 - Preservation
- Indexing and Searching
- User Interface
- Local Language Support

System software stack for implementation

- Red Hat Enterprise Linux ES 4.0
- DSpace 1.4.1
 - Apache, Tomcat
 - Java 1.3/1.4, JSP 1.2, Servlet 2.3

- PostgreSQL 8.2.x, JDBC (rdbms)
- CNRI Handle System 5 (persistent ids)
- Lucene 1.2 (index/search)
- Log4j (logging)
- Implement OAI-PMH and Handle System

DSpace data model

See Figures 1–3.

DSpace information model

See Figure 4.

IVH system model

See Figure 5.

IVH METADATA MODEL

The metadata specific to the IVH project is given below. It has been mapped onto the Dublin core metadata³ and incorporated into DSpace. The metadata was classified into categories such as coverage, identifier and collectors, which are basically elements of the Dublin core standard, and the qualifiers have been coined anew as per the requirements (see Table 1).

CUSTOMIZATION PITFALLS

On Linux

- There are two different directory structures in DSpace. One is the DSpace source and the other is the installation directory the DSpace source is used to build the war file and in fact the installation directory is created using the ant target fresh install.
- To run the fresh install target, the PostgreSQL¹ database should be present and dspace.cfg should be changed in the configuration folder of the source directory. After the fresh install, the installation directory, as specified in the dspace.cfg of the DSpace source directory, is created, and hereafter all configuration changes must be made to the dspace.cfg (which is copied into this folder) in the configuration folder of the installation directory, and the example of the installation directory would be /dspace, which is the default.
- Once fresh install has been run successfully, you need to create an administrator. This is done by going to the bin directory of an installation folder such as /dspace/bin and running the ‘create administrator’ command.

Figure 1: Data model in the form of an object diagram.

- To import metadata and items into the system using bulk import, the files and the dublin core.xml should be included in two levels of nested folders with a separate folder for each item.
- To create thumbnails, you will have to make an entry in the crontab. Entry in the crontab should not be wrong and the cron will automatically start – you don't need to start it again.
- To modify jsps, you need to understand the structure of the page layout. The page is composed of five custom tags, and there are

intelligent defaults to it. The DSpace layout, sidebar, in the DSpace layout, you can turn off the navigation bar and other things such as navlink and toplinek.

- When adding a new jsp, make sure it is being called through a servlet, and that there is an appropriate mapping for the servlet in the web.xml.

On Windows

- You need to give the path of the installation directory in the dos style, as in c:/dspace.

Figure 2: IVH entity relationship.

- You need to run the commands, such as 'create administrator', using dsrun.bat.
- Run the 'bulk import' using dsrun.bat.

INDEX PAGE

This is the index page of the BotanyISpace where you can find general information about IVH, Botanical Survey of India, NIC and the contact details for the Botanical Survey of India. The image on the right-hand side has a rollover effect, and it changes as the mouse is moved into or out of the image. At the bottom of the page is a link that reads 'enter Ispace'. Clicking on this hyperlink takes you to the homepage of BotanyISpace (see Figure 6).

Homepage

The homepage of the installation is the place where the actual functionality can be found, and it is a sitemap for the rest of the application,

Figure 3: IVH entities, preservation stream and meta-data relationship.

Figure 4: IVH information model.

Figure 5: IVH system model architecture.

although quite a lot of the functionality is hidden inside MyBotanyISpace (see Figure 7).

The homepage is divided into five sections: the header, the footer, the sidebar, the navigation bar and the body. This is achieved through a jsp tag that is the layout tag, which embeds the different tags for the other elements. As all the pages have been written inside the layout tag, they can be configured to have any of the elements mentioned above.

In the navigation bar, there are several links, which cater to different functionalities such as browsing, searching and signing on to various sub-functionalities.

Advanced search

This page is displayed upon clicking the 'advanced search' hyperlink in the navigation bar (see Figure 8).

Table 1: IVH metadata model

<i>Metadata description</i>	<i>Fully qualified Dublin core</i>
Accession number	dc.identifier.accessionnumber
Barcode	dc.identifier.barcode
Plant part	dc.plant.part
Prep_code	dc.plant.prepcode
type_status	dc.plant.typestatus
type_qualifier	dc.plant.typequalifier
Date created	dc.date.created
Entered by	dc.contributor.enteredby
Modified by	dc.contributor.modifiedby
Phenology	dc.plant.phenology
Collector	dc.contributor.collector
Collectionnumber	dc.identifier.collectionnumber
CollectedOn	dc.date.collected
Donors	dc.contributor.donor
Donated on	dc.date.donated
Determiner	dc.contributor.determiner
Determinate date	dc.date.determinate
Family	dc.taxonomy.family
Genus	dc.taxonomy.genus
Species	dc.taxonomy.species
Author	dc.contributor.author
Intraspecific rank	dc.taxonomy.intraspecificrank
Intraspecific name	dc.taxonomy.intraspecificname
Intraspecific author	dc.contributor.intraspecificauthor
Determination note	dc.description.determinationnote
Country	dc.coverage.country
Herbarium region	dc.coverage.region
Latitude	dc.coverage.latitude
Longitude	dc.coverage.longitude
Max altitude range	dc.coverage.maxaltituderange
Min altitude range	dc.coverage.minaltituderange
State	dc.coverage.state
District	dc.coverage.district
Precise location	dc.coverage.precise
Locality description	dc.coverage.description
Habitat	dc.plant.habitat
Plant description	dc.description.plant
Habitat	dc.description.habitat
Uses	dc.description.uses
Comments	dc.description.comments

The page provides advanced search features such as searching the entire BotanyISpace repository, or searching a particular collection in BotanyISpace. The search provides conjunction operators that allow three parameters to be searched at a time. They can be associated with different operators such as the common ones AND, OR and so on. The fields on which the search can be performed are grouped into taxonomy, contributors, plant details, identifiers, and so on, which requires prior knowledge of the terms. Taxonomy is a group that includes fields such as family, genus, species and intraspecific rank. Plant details are a group that

includes other plant details such as fringing and holotype.

Advanced search result

The results shown on the page are just a summary, and the hyperlinks in this page on being clicked give the actual metadata list and make the object available for dissemination (see Figure 9).

Metadata list

Upon clicking 'file hyperlink' in the metadata list, the image is displayed, and if an entry has been made for the media filter command in the crontab file in a proper manner then a

Figure 6: BotanyISpace index page.

Figure 7: BotanyISpace home page.

thumbnail will appear in the metadata list page. This page also provides for the dissemination function of the Application, as the user can download the file or the object from here (see Figure 10).

BROWSE

Browse by accession number

This screen is displayed on clicking the hyperlink in the navigation bar. The page

Figure 8: BotanyISpace advanced search.

Figure 9: BotanyISpace advanced search results.

Figure 10: List of metadata.

Figure 11: Browse by accession number.

Figure 12: Browse by collector results.

Figure 13: Results for selected collector.

displays a list of items according to accession number, as well as the collector and the family name. If there are many results, then you can jump to any item within the alphabet (see Figure 11).

Browse by collector

This page is displayed by clicking on the 'By Collector' hyperlink in the navigation bar, and it shows a list of all the collectors in BotanyISpace. Upon clicking on the collectors,

the list of items for the collector are displayed (see Figure 12).

Items for collector

This page is displayed on clicking the results in the previous page, and it lists all the items for the collector on which the user has clicked. There is an option to sort by accession number, where the default behavior is to sort by date (see Figure 13).

Metadata list page

This page is the same as that displayed when the user clicks on 'search results'. It displays the metadata of the item, as well as the filename, description, size, format and the thumbnail of the object that has been saved in the repository (see Figure 14).

Browse by family name

This page is displayed by clicking 'browse by family name' in the navigation bar, and it displays the list of all family names in BotanyISpace. The user can go to any alphabetical element in the result set by clicking on the desired letter in the 'jump to' option above. Or he can enter the first few letters of the family name and hit the 'go' button and the

result list will be redirected to the desired name (see Figure 15).

MyBotanyISpace

A login screen is displayed upon clicking the MyBotanyISpace hyperlink in the navigation bar, and to log in you require a valid username and password, which is nothing but a registered e-person with the system. We will see the creation of the e-person later on. At the time of the installation, an administrator is created, and this will be used to log in to the system at the first instance to create other e-people (see Figure 16).

Tasks in pool and submission status

After logging in to MyBotanyISpace, you will find the tasks that have been created and need approval from you. You need to click on the 'take task' button and follow the process thereafter (see Figure 17).

ADMINISTRATION TOOLS

The Administration module consists of managing epeople, items, policies, groups, metadata registry,

Figure 14: Metadata list for the browse results.

Figure 15: Browse by family name.

Figure 16: Login screen.

Bit-stream format registry, workflow, authorization, supervisors and statistics. Given below are the web pages that will illustrate the use and utility of the tools and how they are relevant to the application (see Figure 18).

Administration tools provide the user with administrative rights to modify core functionality of IVH Space, such as adding metadata, editing metadata, adding supported formats to the system, modifying and adding workflow items, and modifying authorization policies.

Ingest workflow

The ingest workflow starts after you log in to My IVH Space and click on the 'herbaria and collections' hyperlink. You get a list of herbaria and collections, and when you click on the collection within a herbarium you will get the collection home page. On the collection home page click on the 'submit to collection' button. Then the ingest workflow starts. What subsequently follows is a set of pages that constitute an ingest workflow (see Figure 19).

Figure 17: Tasks in pool.

Figure 18: Administration tools.

Describe metadata step 1

This is the first step to describe the metadata, and it asks a question as to whether the item to be ingested consists of more than one file. Click on the 'next' button to navigate to the next screen (see Figure 20).

Describe metadata step 2

This step describes the following metadata about the plant specimen: Barcode, Accession Number, Plant part, Prep Code, Type Status and Type Qualifier (see Figure 21). This is basically a set of Metadata about the plant

Figure 19: Collection home page.

Figure 20: Describe metadata step 1.

specimen's physical properties and also some date specification. You have to be an informed user to enter all these metadata, and the

authorization policy can decide on whom to give the rights for the submission of documents.

Figure 21: Describe metadata step 2.

Figure 22: Edit metadata step 3.

Edit metadata step 3

This step describes the taxonomy of the plant: family name, genus, species and so on (see Figure 22).

Edit metadata step 4

This step describes the collection details (see Figure 23).

Edit metadata step 5

This step describes the coverage details, such as the location, region and country where the plant is found and the altitude at which the plant is found (see Figure 24).

Describe metadata step 6

This page describes the notes associated with the plant specimen (see Figure 25).

UPLOAD FILE

In this step, you can upload the plant image that you want to associate with the metadata (see Figure 26).

Verify step

In this step, the submitter can verify whether or not the metadata entered is correct. If

modification is needed, it can be corrected. Even if the file that has been uploaded needs to be changed, this can also be done (see Figure 27).

Once it has been verified that the metadata and the file to be uploaded are correct, the user can move to the next step, which is the final step.

Accept license step

Once you grant the license, the file will be uploaded in the system and the metadata will be bound to the file, and hence the ingest process is complete (see Figure 28).

SUMMARY

IVH Space as knowledge archival system

IVH Space a customized version of DSpace 1.4.2, and serves as an archival storage system. The archival process consists of a chain of sub-processes and a workflow system that enables the review of the items being ingested, while at the same time providing checks for the submission manager to validate the metadata before it is finally ingested into the system. Pertinent to tape storage, there is a utility called Storage Resource

Figure 23: Edit metadata step 4.

IVHSpace: Describe this Item - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://localhost:8080/dspace/submit#null

Describe Describe Describe Describe Describe Describe Upload Verify Licen

Submit: Describe this Item

Please fill further information about this submission below. ([More Help...](#))

Select the country for the specimen. To select more than one value in the list, you may have to hold down the 'CTRL' or 'Shift' key.

Country

Enter the Herbarium region in which the plant thrives

Herbarium Region

Enter the longitude

Longitude

Enter the maximum altitude at which the plant is found

Max Altitude

Done

Gmail - I... [root@s... IVHSpac... dspaced... dsp - Op... bispaces...

Figure 24: Edit metadata step 5.

IVHSpace: Describe this Item - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://localhost:8080/dspace/submit#null

Describe Describe Describe Describe Describe Describe Upload Verify Licen

Submit: Describe this Item

Please fill further information about this submission below. ([More Help...](#))

Enter a short plant description

Plant Description

Enter the Habitat of the plant

Habitat

Enter the Uses of the plant

Uses

Done

Gmail - I... [root@s... IVHSpac... dspaced... dsp - Op... bispaces...

Figure 25: Edit metadata step 6.

Broker, which can be configured to transfer the files from the Server hard disk to the tape library. The Ingest workflow is such that

functionality is not compromised. The submission workflow is customizable to a high degree, and additional pages and additional

Figure 26: Upload file step.

Figure 27: Verify submission step.

metadata can be added (which of course requires the metadata to be first registered with the system). Once the submission process is

complete, the metadata and the document content can be reaped through an efficient and comprehensive search, which is available to the

Figure 28: Grant license step.

general user (in DSpace terms, an anonymous user). Lucene search engine, an open source search engine provided by Apache, is used to do the search. It requires that the system be indexed beforehand, which can be done through a command that can be run as a cron daemon in the UNIX environment. As the DSpace system is indexed, it becomes available for exploiting search. There are other additional features such as a bulk import facility and metadata harvesting. Statistics can also be generated, which can be

useful for auditing purposes. The statistics package is also fully configurable. The configuration of DSpace is done through the `dspace.cfg` file, which is a text file with key value pairs.

REFERENCES

- 1 <http://www.postgresql.org/>.
- 2 <http://www.dspace.org/>.
- 3 <http://dublincore.org/>.
- 4 <http://envfor.nic.in/bsi/>.