Digital Archiving and Management at National Informatics Centre

Digital objects are being created by governments, private corporations, authors, publishers, librarians, museum curators etc. in the form of text documents, photographs, images, audio and video footage, maps and more. Specifically, the massive volumes of paper generated by government offices can be pulled out of files and digitized. Metadata can be utilized to easily categorize digitized records of government offices. This would facilitate access to information buried deep in files, making it available at the click of a mouse.

Nandita Chaudhri
Senior Technical Director
nandita@nic.in

Surinder Kumar
Technical Director
suri@nic.in

The rapid surge in the generation and dissemination of digital information, in various formats, is a unique phenomenon of our times. Huge volumes of information, in digital form, can be stored more compactly and accessed with greater speed and ease than on paper. In a networked environment, these attributes of digital access take on mindboggling dimensions. On the other hand, digital information is fragile unlike information on traditional media such as paper. It is vulnerable to corruption, to alteration without detection and to obsolescence of storage and access technologies. As more and more information is either 'born digital' or digitized from its analog form, a special focus is required to archive and manage it efficiently and effectively so that it is available over longer periods of time and is not lost to posterity. Accordingly, the science of Digital Archiving and Management has evolved rapidly over the last decade.

About DSpace:

Based on careful evaluation of the different technology options, the Digital Archiving and Management (DAM) group of NIC has identified Open Source technology, to provide solutions to other government organizations/agencies. DSpace is being used currently, which is a groundbreaking Open Source digital repository software tool which captures, stores, indexes, preserves and redistributes content in digital formats.

- System Architecture: The main code of DSpace is implemented in Java, and runs on any UNIX-like system. It makes use of several third-party open source systems: PostgreSQL, an open source relational database system; Jakarta Tomcat Java Servlet container and Apache HTTPD server, for optional SSL and X.509 certificate support

- Search and Browse: DSpace uses the Lucene Search engine. It provides fielded searching, stop words, stemming. Its API allows indexing new content, regenerating the index, and performing searches on the entire corpus, a community, or collection. Another important mechanism for document discovery in DSpace is the 'browse'.

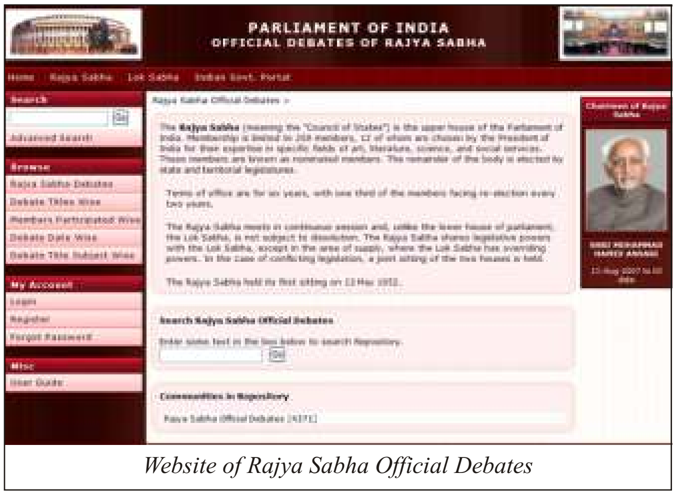
- Content Organization: Content, at the highest level, in DSpace, is organized into communities. These correspond to organizational bodies in an institution, such as departments, labs, research centers or schools. A community is organized into collections of logically-related material. For example, a technical report series might be a collection. An item is an “archival atom”; that is, a grouping of content and metadata that it makes sense to archive as a single unit. Each item has one Dublin Core metadata record. There is a facility for Bitstream storage, Persistent Naming and Personal Workspace. Also available is a feature of an editorial review process which provides three tiers workflow in submission of an article.

- Look and Feel: The latest version of DSpace has an XML based UI, as well, known as Manakin. Manakin is the second version of DSpace XML UI that uses SAX & the Cocoon framework to enable communities and collections to establish a unique look and feel that is distinct from the default installation of DSpace.

Major projects

Some major projects undertaken, by the group, are

- Digitization of Rajya Sabha Debates: The nation's most august body, the Rajya Sabha, or the Upper House of the Parliament of India, during its sessions, witnesses Questions and Answers and Debates in English, Hindi and Urdu. A record of these is presently available in print format. Towards the goal of building a digital repository of these proceedings, the work for digitization has been outsourced and NIC has undertaken the task of archiving the digitized information using DSpace. This repository would provide invaluable reference material for legislators and the general public alike.

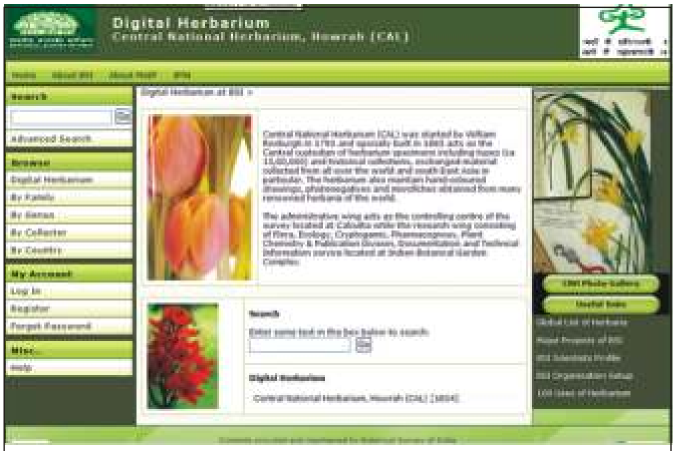


Website of Rajya Sabha Official Debates

- Knowledge Repository of Inter-State Council Secretariat: ISCS@Digital Repository has been developed using DSpace, and can be accessed over the Local Area Network of the Council. The repository contains the Report on Centre State Relations of the Sarkaria Commission. The Action Taken Report of one chapter of this report has presently been included in the repository as a pilot project.

Indian Virtual Herbarium (IVH) & Digital Herbaria (DH)

The conservation of biological diversity is a critical necessity for the maintenance of the Biosphere in a state supportive of human survival on Earth. The Botanical Survey of India (BSI) under the Ministry of Environment and Forests (MoEF), acts as the custodian of the country's floral wealth comprising about 45,000 plants spread over the entire country. NIC has developed a solution for implementing a Digital Herbarium (DH) comprising an Imaging Studio, as well as complete repository of high resolution images of this flora, with search facility, at each of BSI's 14 Herbaria, for in-house access. Also, an Indian Virtual Herbarium (IVH) is being developed at a central location with high speed Internet connectivity.



Webpage of Digital Herbarium at BSI

The DAM group has also provided a solution for the Administration branch of UPSC, for digital archiving of its records. Currently under consideration, are digitization projects for documents belonging to the Ministry of Environment and confidential records of the Anti-Crime Bureau of Hyderabad. *i*

For further information, contact:

National Informatics Centre
Digital Archiving and Management Group
damgrp-list@lsmgr.nic.in
Tel: 91-011-24305520/ 24364265