IN13A-1827 AGU, Fall 2015

Curating and Preserving the Big Canopy Database System: An Active Curation Approach using SEAD

James D. Myers[1], Judith B. Cushing[2], Peter Lynn[2], Noah Weiner[2], Anna Ovchinnikova[1], Nalini Nadkarni[3], Anne McIntosh[4]

[1] Inter-university Consortium for Political and Social Research, University of Michigan, Ann Arbor, MI, myersjd@umich.edu
[2] The Evergreen State College, Olympia, WA, judyc@evergreen.edu
[3] University of Utah, Salt Lake City, Utah
4 University of Alberta, CANADA

Preserving At-Risk Scientific Data and Cyberinfrastructure

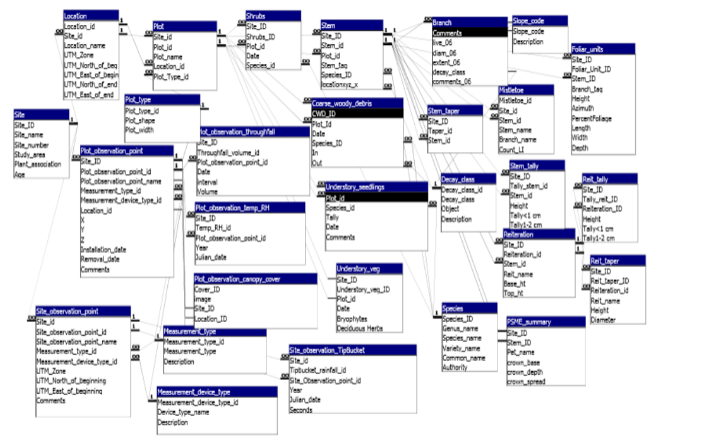
Modern research is increasingly dependent upon highly heterogeneous data and on the associated cyberinfrastructure developed to organize, analyze, and visualize that data. However, due to the complexity and custom nature of such combined data-software systems, it can be very challenging to curate and preserve them for the long term at reasonable cost and in a way that retains their scientific value. In this presentation, we describe how this challenge is being met in preserving an at-risk collection of data and tools -- Canopy Science Data and Applications (Canopy DataBank, and the databases created using this software, the Big Canopy Database (BCD), and CanopyView) – using an agile approach and leveraging the Sustainable Environment – Actionable Data (SEAD) DataNet project's hosted data services. The Canopy DB Project applications were developed over more than a decade and have been used to record irreplaceable observational data from field research on a broad range of forests.

Canopy Cyberinfrastructure

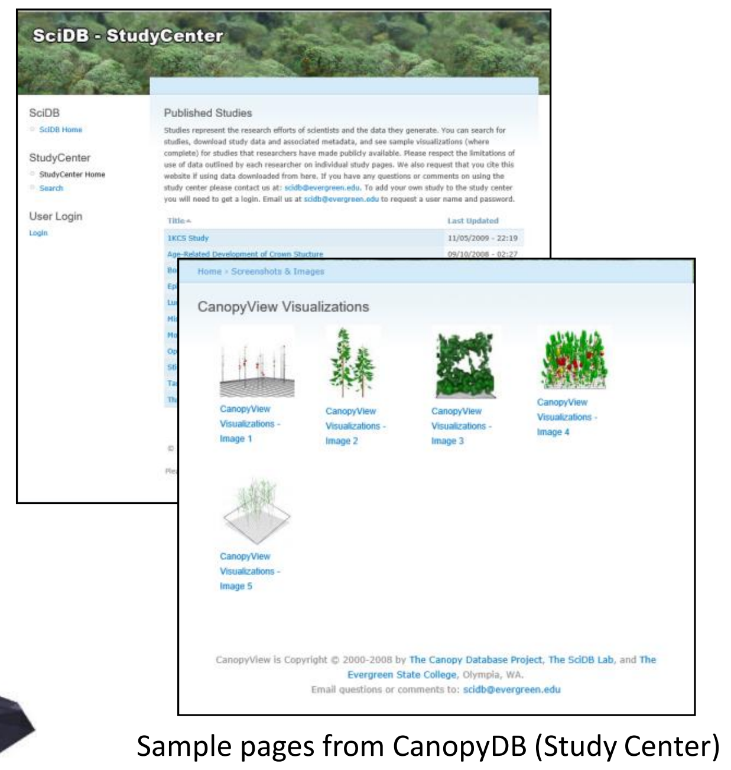
The CanopyDB applications were developed, with support from the U.S. National Science Foundation, over more than a decade at The Evergreen State College to address the needs of forest canopy researchers. CanopyDB is an early yet sophisticated exemplar of the type of cyberinfrastructure that has become common in biological research and science in general, with multiple relational databases for different experiments, a custom database generation tool used to create the databases, an image repository, a bibliographic and research reference tool, and desktop and web tools to visualize the data.

Components:

- **Databank Database Generator:** Graphical tool to generate ecological databases from standard and custom database templates. Generated database includes data entry forms, data dictionary, and Ecological Metadata Language (EML) documents.
- **CanopyView:** An interactive visualization tool designed to view tree structure, canopy coverage, and other data stored in Databank-generated databases.
- **Project Website:** The project maintained an extensive website providing access to the cyberinfrastructure and **a catalog of data created by projects** using it. The website contains extensive metadata and documentation and software and data download links. It also includes a general image catalog capability managing **collections of photos and visualization images**.



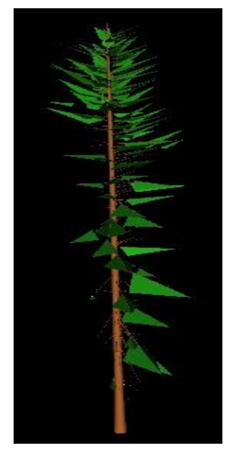
Canopy DataBank 1kcs database E-R diagram


SciDB - StudyCenter
SciDB
SciDB Home
StudyCenter
StudyCenter Home
Search
User Login
Login
Published Studies
Title
Last Updated
1KCS Study
11/05/2009 - 22:19
Age-Related Development of Crown Structure
09/10/2008 - 02:27
Home › Screenshots & Images
CanopyView Visualizations
CanopyView Visualizations - Image 1
CanopyView Visualizations - Image 2
CanopyView Visualizations - Image 3
CanopyView Visualizations - Image 4
CanopyView Visualizations - Image 5

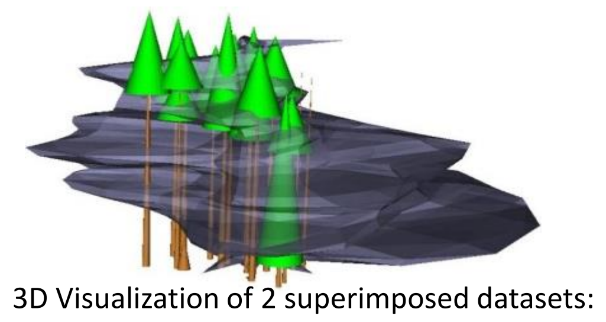
Studies represent the research efforts of scientists and the data they generate. You can search for studies, download study data and associated metadata, and see sample visualizations (where complete) for studies that researchers have made publicly available. Please respect the limitations of use of data outlined by each researcher on individual study pages. We also request that you cite this website if using data downloaded from here. If you have any questions or comments on using the study center please contact us at: scidb@evergreen.edu. To add your own study to the study center you will need to get a login. Email us at scidb@evergreen.edu to request a user name and password.

CanopyView is Copyright © 2000-2008 by The Canopy Database Project, The SciDB Lab, and The Evergreen State College, Olympia, WA.
Email questions or comments to: scidb@evergreen.edu

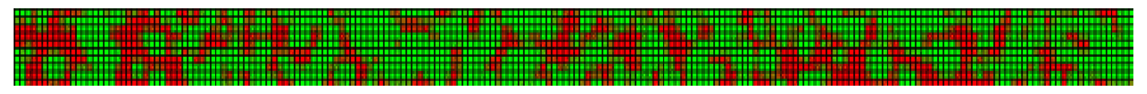
Sample pages from CanopyDB (Study Center) and Visualization Repository web site.



Screenshot of visualization depicting 3D tree structure.



3D Visualization of 2 superimposed datasets: cluster of trees and understory open space.



A simple 2D visualization showing amount of overhead canopy cover for a study plot. Green indicates covered areas; red is "open" space.

Canopy Data

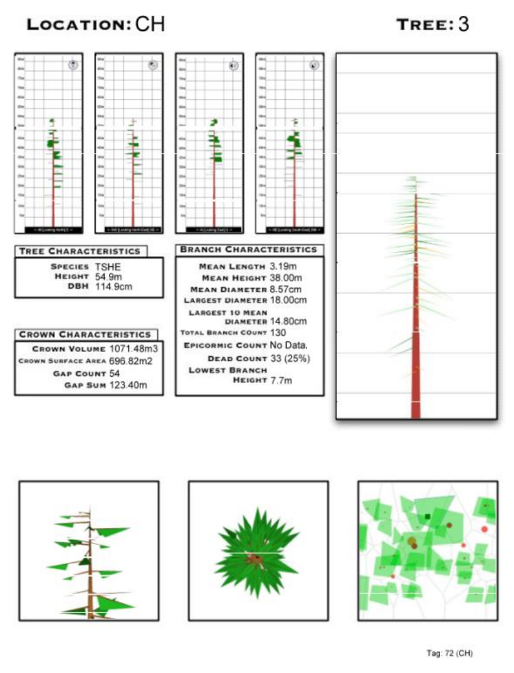
The Databank database generator was used to create structured data repositories for a number of forest studies. The most extensive of these was a major study of forest structure dubbed the Thousand Year Chronosequence, 1kcs. Objectives of this study were to characterize the composition, density, surface area, biomass, and spatial distribution of trees, saplings, and understory vegetation in a chronosequence of eight Douglas-fir/western hemlock stands ranging in age from 50 to approximately 950 years, all located in the Western Washington Cascades: Wind River Experimental Forest, Carson, WA; Mt. Rainier National Park; and Cedar Flats Research Natural Area. The study took place from September 1, 1999 to September 1, 2006.

These data were used in a subsequent project, the "Evergreen-DNR (Department of Natural Resources) Leave Tree Project". The project goal was to help forest managers of Washington State Public Lands use results from forest canopy research to help determine which trees to leave when harvesting a stand of trees, exploring how tree features such as: complex structure, broken tops, large branches, crown gap, and continuous crown could be related to policy goals such as leaving "trees that increase wildlife habitat". As part of this work, CanopyView was used to generate visualizations of 100 trees for which detailed structure had been recorded.

Although the level of documentation varies by study, the project website includes a wide range of documentation of the efforts including descriptive text, schema diagrams, data dictionaries, publications and presentations, executable software, manuals, photo galleries, guides describing how to navigate to field sites, and descriptions of experimental procedures and calibrations.

Researcher	Database
Bob VanPelt	Thousand Year Chronosequence
Hiroaki Ishii	Age-Related Development of Crown Stucture
Roman Dial	Borneo Insect Biomass and Count
Betsy Lyons	Epiphytes and Hemlocks
Eda Menendez	Luquillo Canopy Plot Visualization
David Shaw	Mistletoe and Hemlocks
Traci Sanderson	Monteverde, Epiphyte Changes Over Time
Roman Dial	Open Space in Canopy Structure
Akiro Sumida	Stick structure of Japanese chestnut
Yoav Bar-Ness	Tasmanian Eucalyptus obliqua: Crown Structure and Arthropod Biodiversity
Geoff Parker	Three-Dimensional Canopy Structure

Table 1: Databases created using DataBank, and stored in the CanopyDB repository (aka Study Center).



A sample page from the DNR 1kcs Catalog

SEAD Data Services

SEAD Sustainable Environment Actionable Data
University of Michigan · Illinois · Indiana University Bloomington

SEAD, created through the U.S. National Science Foundation's DataNet program, provides secure hosted project spaces in which teams can incrementally develop data collections in preparation for publication and long-term preservation. SEAD's iterative model, in which data can be added, organized, and annotated over time, using an evolvable set of community and custom vocabularies, and in which data can be published at any desired level of granularity to a range of institutional and community repositories, was designed to support curation in parallel with research activities. However, it is proving to also be well suited to the reconstruction of records from distributes sources at the end of a project's active phase.

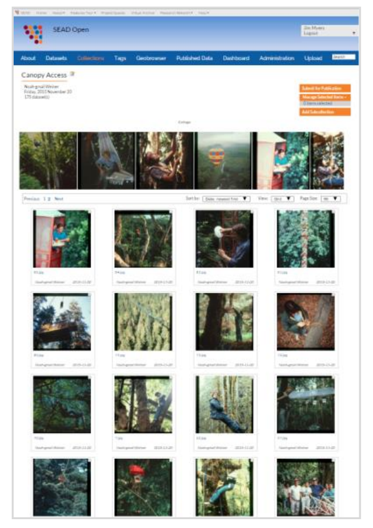
SEAD Project Spaces are hosted web applications with interactive user interfaces and RESTful programming interfaces. SEAD also provides tools for bulk uploads (100-100K+ files) , batch submission of metadata, and interacting with content from applications such as R. SEAD also provides services for publication that can match content to appropriate repositories and manage publication of multiple versions of content over time. Publication in SEAD includes generation of a persistent global identifier for use in data citation, and, for large collections, packaging of data into Open Archives Initiative Object Reuse and Exchange (OAI-ORE) and Library of Congress BagIT compliant containers that simplify long-term management.

SEAD
Sustainable Environment Actionable Data
SEAD-DATA.NET
PROJECT SPACES
LIVE, COLLABORATIVE SPACES TO SHARE, ORGANIZE, AND ANNOTATE DATA
DATA CURATION & PRESERVATION
STAGING AREA WITH MATCHMAKER TO GUIDE YOUR DATA TO A TRUSTED REPOSITORY
DATA MANAGEMENT
DATA SHARING AND ARCHIVING DATA PLAN SUPPORT
PUBLISHING YOUR DATA
MAKE YOUR DATA CITABLE WITH THE PUSH OF A BUTTON
Open source, lightweight tools for data-intensive sustainability and social science research to enable sophisticated management of heterogeneous data and to dramatically lower the effort required to curate and preserve data for the long term.
SEAD is funded by the National Science Foundation under cooperative agreement #OCI0940824.
UNIVERSITY OF MICHIGAN
ILLINOIS
INDIANA UNIVERSITY

Active Curation to Rebuild Scientific Records

The effort to preserve the cyberinfrastructure of the Canopy DB projects and the data collected using it began as servers housing it were being decommissioned and after the development and field teams had dispersed to other institutions. The following items outlines the ongoing process we are using to assemble and prepare the records for publication and describes some of the challenges being faced:

- **Assembly** – CanopyDB software and data were primarily made available via the project's website, but software and relevant documentation existed on additional servers and in personal archives. We used SEAD's bulk upload capability to ingest 4597 files from two servers that housed the main website, making that content available for viewing by our distributed team. As the scope of that material was examined, it was used to guide exploration for material on additional servers (one requiring a RAID array rebuild) and personal computers.
- **Deduplication** – the raw uploads included hundreds of duplicate files and many instances where files with different names/locations had the same content. We used SEAD's automatic generation of SHA1 hash signatures for files and interactive SPARQL query capability to identify these duplicates and it's tagging and comment capabilities to alert team members. We also deployed a 'SeenInSEAD' service that can check local files to see if a copy already exists in SEAD (without requiring the files to be uploaded).
- **Exploration** – As team members reviewed the growing collection, we have used tags, comments, and formal metadata to add descriptive information and to categorize content, e.g. "data dictionary", "commercial software", "image gallery". SEAD's built-in ability to preview many types of files (text, pdf, csv, images, movies, ppt, etc.) made it easy to explore. In this sense, SEAD also replaced/obviated the need for commercial image gallery software (a version which is now known to have security issues) that had been part of the original CanopyDB website.
- **(Re)Organization and Enhancement**– The structure of materials on disk were dictated by web server requirements and other project requirements rather than scientific concerns. Web links were used to add a second layer of organization. Metadata was spread across file naming conventions, web page content, documents, and, for the image gallery and canopy databases, in database tables. In SEAD, we have the ability to directly annotate data files with formal (RDF) metadata and make associations between files and other files and/or external resources (e.g. people, publications , and software source code described/managed in third-party repositories). Given the richness of the information available and the scope of the materials, we have elected to experiment manually within SEAD to develop a reasonable practice and to then explore the use of automated scripts to enhance the overall corpus.
- **Publication** – SEAD's publication process generates a Digital Object Identifier (DOI) and serializes all content and metadata into a single standards-based compressed package. Given the ability to publish multiple versions, we intent to publish an initial content-complete version and then create updates as we enhance the scientific reusability with further effort. As a proof-of-concept, we have run an in-progress version of the CanopyDB materials through SEAD's publication process, resulting in a temporary DOI (http://dx.doi.org/doi:10.5072/FK2QR4TC0R) that resolves to a landing page at https://sead-test.ncsa.illinois.edu/ndsrepository/landing.html#tag%3Asead-data.net%2C2015%3ARO_EB7slrT2iC7pjysQR2QuaQ.



The CanopyDB Image Gallery, recreated within the team's project space in SEAD

Conclusions and Future Work

Using SEAD, our cross-project team is incrementally ingesting CanopyDB components (images, datasets, software source code, documentation, executables, and virtualized services) and is iteratively defining and extending the metadata and relationships needed to document them. However, we also recognize that publication of the files from the CanopyDB project is not sufficient to enable future researchers to interact with its databases and visualization tools. We therefore anticipate further work to explore options such as extracting the databases into database-agnostic formats, creating and preserving virtual machine images or containers, extracting materialized views as files stored and annotated in SEAD, or, most ambitiously, extending SEAD's preview capabilities using facilities such as the National Data Service Labs environment, to include launching interactive sessions on cloud resources close to the repository.

We strongly believe that preservation of rich, heterogeneous combinations of custom cyberinfrastructure and data, with sufficient fidelity to support re-use, is one of the major challenges to realizing an effective national/global scientific data infrastructure. We hope that our current work will help demonstrate what is possible today and help contribute to the development of best practices for preserving scientific outputs.

References

- Myers, J.D., et.al., *Towards Sustainable Curation and Preservation: The SEAD Project's Data Services Approach*, Proceedings of the IEEE 11th International e-Science Conference, Munich, Germany, 2015, DOI 10.1109/eScience.2015.56
- Cushing, J.B., N. Nadkarni, B Bond, R. Dial. 2003. How trees and forests inform biodiversity and ecosystem informatics. 2003. *Comput. Sci. Eng*. 5, 32 (2003).
- Cushing, J.B., N.M. Nadkarni, M. Finch, A.C.S. Fiala, E. Murphy-Hill, L. Delcambre, and D. Maier. 2007. Component-based end-user database design for ecologists. *Journal of Intelligent Information Systems*. 29(1): 7-24.
- Kopytko, N., J.B. **Cushing**, L. Zeman, N. Stevenson-Molnar, F. Martin, E.W. Keeley. 2009. Making ecology research results useful for resource management: a case study in visual analytics. dg.o '09. _Proceedings of the 10th Annual International Conference on Digital Government Research, pp. 206-215.
- Nadkarni, N.M, A.C.S. McIntosh, J.B. Cushing. 2008. A framework to categorize forest structure concepts. *Forest ecology and Management* (256, 5), pp 872-882.
- Van Pelt, R., and N. M. Nadkarni. 2004. Development of canopy structure in Pseudotsuga menziesii forests in the southern Washington Cascades. Forest Science 50:326-341

Acknowledgements

SEAD is funded by the National Science Foundation under Cooperative Agreement #OCI0940824.
CanopyDB development and use were supported through multiple U.S. National Science Foundation grants including: DBI 04- 17311, CISE 01-31952, BIR 03-19309, 99-75510, 96-3O316, 93-07771.

SEAD gratefully acknowledges all of the partners and participants who have been involved in developing our services framework. The CanopyDB Project likewise gratefully acknowledges its many young developers, including undergraduate students, and ecologists who contributed energy and data to the project.

Visit SEAD at Booth #113!

For More Information

Web: http://sead-data.net/
Twitter: @SEADdatanet
Email: SEADdatanet@umich.edu