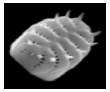
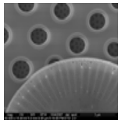
Exploring e-Curation of Diatomscapes via Levels of Digital Curation and The DCC Curation Lifecycle Model

Plato L. Smith II
Florida State University
Tallahassee, FL
psmithii@fsu.edu



Faculty acknowledgement

"I am very pleased that my diatom images are now digitally archived as part of the pilot program for [Florida Digital Archive - FDA] digital preservation. I am honored to be partnering with [FSU Digital Library] and your colleagues [Florida Center for Library Automation - FCLA] on this innovative program. Kindly extend my thanks, on my behalf, to Ms. Motyka and Ms. Caplan for their contributions to the success of this important aspect of our collaboration. I hope it is just the beginning for a long and mutually beneficial partnership between scientists and digital technologists. Thank you again for this exciting news. You made my day". - Dr. A.K.S.K. Prasad, FSU Biological Scientist



Diatomscapes are images of biological silica and consists of diatoms ("microscopic, single-celled plants that thrive in freshwater, saltwater, brackish water and even semi-terrestrial environments" (Prasad, 2005)) and Radiolarian ("amoeboid protozoa that produce intricate mineral skeletons" digital images developed by Florida State University's Department of Biological Science faculty member, Dr. A.K.S.K. Prasad.

Abstract

Clearly formulate the research question

If Levels of Digital Curation and the DCC Curation Lifecycle Model were articulated to a faculty member, would the faculty member be interested in working towards employing digital curation practices for the preservation of his/her digital research data images?

Identify the significant problems in the field of research

An unobtrusive site visit, informal face-to-face faculty interviews, review sample of images of biological silica images (Diatomscapes), and numerous email correspondences yielded a need for structured descriptive information, introduction of an established metadata standard, creation of discovery & accessibility, assessment of current level of digital curation, need for preservation & storage, and application of the DCC Curation Lifecycle model to the FSU biological science research discipline (the scope of this poster session is limited to the biological science research of Dr. A.K.S.K. Prasad)

Summarize the current knowledge of the problem domain, as well as the state of the art for solutions

➢**Problem:** Descriptive or representative information of images of biological silica do not utilize an established metadata standard
✓**Solution:** introduce and apply the Access to Biological Collections Data (ABCD) metadata standard to the images of biological silica Diatomscapes collections
➢**Problem:** Diatomscapes were offline (created access and facilitated image sharing)
✓**Solution:** Created basic online Diatomscapes collections in Picasa and Flickr
✓**Unexpected outcome –** Scientist's sponsor requested collections URL and increased interface/collaboration with courtesy faculty member
➢**Problem:** Images of biological silica are stored on individual workstations, discs, or networked attached storage without use of an established preservation strategy
✓**Solution:** Preserved Diatomscapes I and II via Florida Center for Library Automation (FCLA) Florida Digital Archive (FDA) and MetaArchive
✓**Faculty benefit –** Dr. A.K.S.K. Prasad's Diatomscapes I and II collections were the first FSU faculty member's research data images to employ either FDA or MetaArchive preservation strategy
✓**Additional benefit –** Two preservation strategies are being used to preserved Dr. Prasad's images of biological silica for current and future use

Clearly present any preliminary research plans and ideas, and the results achieved so far

Build Diatomscapes digital collections in DigiTool; create MARC record in catalog (Aleph); create record in OCLC WorldCat
Currently researching Morphbank use of Darwin Core and ABCD for biological image description for application to Diatomscapes collection
Encourage more Diatomscapes ingestion into Morphbank (bilogical image database)
✓Coordinated faculty meeting of faculty member, faculty member's sponsor, faculty member's colleague & fellow scientist, and principle investigator of Morphbank meeting to discuss a potential future collaboration for grant application submission
✓Organized an "Explore the World of Diatomscapes" exhibit within the libraries which generated interest among scientists, faculty, students (both graduate and undergraduate), campus community and the public
✓Built Diatomscapes collections in Picasa and Flickr
✓Preserved Diatomscapes via FDA and MetaArchive
✓Advertised & marketed faculty member's images of biological silica images via presentations at the 2009 ACRL 14th National Conference, SPARC Digital Repositories Meeting 2008 Innovation Fair, and 4th International Digital Curation conference 2008

Sketch the research methodology that is to be applied

✓Levels 1 , Level 2 ,and Level 3 Digital Curation (Lord, 2003)
✓The DCC Curation Lifecycle Model (DCC, 2007)
❖Boyer's Model of Scholarship (Boyer, 1997) – **NOTE:** future application for metatriangulation

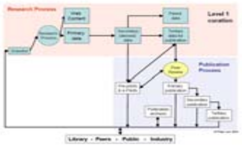
Describe the expected contributions of the applicant to the research area

✓Other FSU scientists and faculty have requested copies of large-format posters, images, poster templates used for the "Explore the World of Diatomscapes"
✓Dr. Prasad's Diatomscapes has been heavily publicized and is getting long-overdue recognition and mass appeal
✓Diatomscapes posters are currently displayed in the FSU Paul M. Dirac Science Library and FSU Libraries Digital Library Center
✓It is expected that the adoption of established metadata standards for description and digital curation practices will become more popular in the future for the FSU biological research discipline as a result of FDA and MetaArchive pilot preservation of Diatomscapes I and II

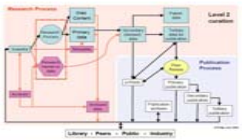
(For technical research) describes how the research is innovative, novel or extends existing approaches to a problem

✓Using the levels of digital curation to assess current digital curation practices of a faculty member's digital objects , The DCC Curation Lifecycle Model as a reference model to guide digital curation practices for digital objects while incorporating Boyer's Model of Scholarship to convey significance of faculty member's research is a novel attempt at metatriangulation ("building theory from multiple paradigms")

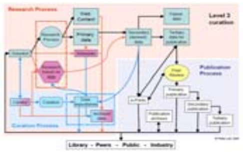
Levels of Digital Curation



Traditional academic information flow – Level 1 Curation



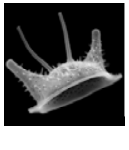
Information flow with data archiving – Level 2 Curation



Information flow with data curation – Level 3 Curation

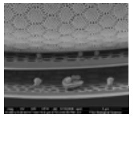
Mapping to Diatomscapes

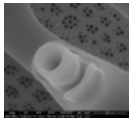
- •Diatomscapes images on discs; PC; offline
- •A few images in Morphbank (< 14)
- •A few images published in journals
- •Images identified & selected for collection building
- •No current digital curation efforts
- •Images at risk of lost

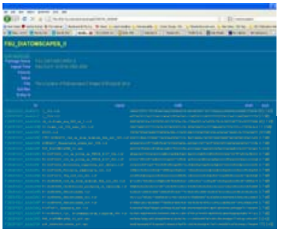


- •Diatomscapes images metadata organized (gathered/updated) by scientist
- •Digital librarian began to work with scientist to build digital collection
- •Diatomscapes images transformed into JPEGS & uploaded to Picasa/Flickr (online/restricted)
- •Images moved from discs to networked attached storage
- •Developed Diatomscapes exhibit

- •Diatomscapes I and II images checksum and METS packages created for FTP to FDA
- •LOCKSS Manifest pages, XML plugin, conspectus database, and preservation cache completed for MetaArchive preservation
- •Diatomscapes I and II images are currently being preserved via FDA and MetaArchive

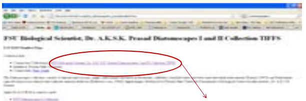






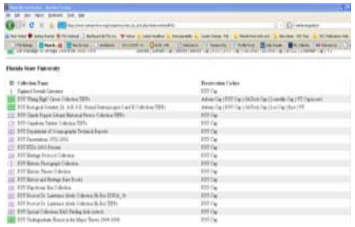
Florida Digital Archive (FDA) - Package Name; Ingest Time; Title; id; name; md5; sha1; size (online collection preservation statistics)

LOCKSS Manifest page – grants permission for LOCKSS to crawl and harvest archival units represented as URLs (hyperlinks)



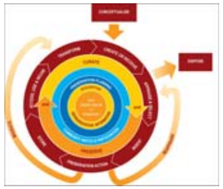
Points to RDF Descriptive Data from conspectus database entry

Collection Description Data Editor – describes digital collection in the MetaArchive Conspectus Database



Collections with green IDS are marked to be preserved (their archival units are part of the title database – Auburn Cap | FSU Cap | GA Teach Cap | Lou Cap | Rice | VT preservation caches

The DCC Curation Lifecycle Model



The DCC Curation Lifecycle Model - The DCC Curation Lifecycle Model provides a graphical high level overview of the stages required for successful curation and preservation of data from initial conceptualization or receipt. The model can be used to plan activities within an organization or consortium to ensure that all necessary stages are undertaken, each in the correct sequence. The model enables granular functionality to be mapped against it; to define roles and responsibilities, and build a framework of standards and technologies to implement. It can help with the process of identifying additional steps which may be required, or action which are not required by certain situations or disciplines, and ensuring that processes and policies are adequately documented.

➢**Data (Digital Objects or Databases)**

Digital object – Simple Digital Objects are discrete digital items, such as textual files or sound files, along with their related identifiers and metadata.

✓**Diatomscapes –** 38 images of biological silica were donated via compact discs by FSU Biological Scientist, Dr. A.K. S.K. Prasad to be used as part of this project. The images of biological silica represent diatoms and radiolarian gathered from 2004 – Present mainly from the southeastern most part of the US. The images were in tagged image file format (TIFF). However, the stipulation was given that the images are not published online until some species name new to science have been published via traditional publishing venues. Permission to share images is only based on personal requests (limited/restricted access). The researcher's view of enacting limited access and the institution's view of what data to preserve are not new issues as the following excerpt from the Research Information Network (RIN) suggests. "Researchers funders and institutions need to take full account of the different kinds of data that researchers create and collect in the course of their research, and of the significant variations in researchers' attitudes, behaviors and needs, and to make clear the categories of data that they wish to see preserved and shared with others in each case" (Griffiths, 2008).

Databases – Complex Digital Objects are discrete objects, made by combining a number of other digital objects

➢**Full Lifecycle Actions**

Description and representation information – assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards to ensure adequate description and control over the long term.

✓**Diatomscapes –** Diatomscapes lacked complete metadata and only consisted of basic metadata elements which could be easily mapped to Dublin Core for simple metadata record to be used in an institutional repository such as DigiTool. Those basic metadata elements included scientific name, collection site, date of collection, ecological preference, image id (or unique identifier), and microscopic magnification. The paper proposes gathering more technical metadata for the images of biological silica and use Access to Biological Collections Data (ABCD) as a metadata content standard and Metadata Encoding Transmission Standard (METS) as a digital content standard. Currently, there are no appropriate standards or adequate description of the images of biological silica and thus "…newly created data sets continue to enter the same trajectory of degradation and loss that has been and is now experienced by legacy data set" (Kintigh, 2006, p. 572).

Preservation Planning – Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.

✓**Diatomscapes –** Diatomscapes was used as a Florida State University demo test collection for preservation using Dark Archive in the Sunshine State (DAITSS) open source software as part of the Florida Digital Archive (FDA) at Florida Center of Library Automation (FCLA). The preservation planning of Diatomscapes included creation of METS descriptor file packet which included technical metadata and MD5 check for each digital object within a METS package, online preservation statistics reporting, and ingest report. At the time of the project, FSU Libraries did not have an account established with FDA for the preservation of FSU digital content. The preservation planning of Diatomscapes via FDA includes the management, preservation, and online reporting statistics such as file size, MD5/Sha1, # of files, and file names of preserved content. "In the future, a scholar or researcher will want to know that a digital object is trusted – that is authentic and reliable "(Jantz, 2008).

Community Watch and Participation – Maintain a watch on appropriate community activities, and participate in the development of share standards, tools and suitable software.

✓**Diatomscapes –** The researcher has been introduced to Marine Metadata Interoperability (MMI) and Access to Biological Collection Data (ABCD) as a result of this project. The ABCD is a content standard in use with the Global Biodiversity Information Facility (GBIF) and Biological Collection Access Service for Europe (BioCASE) and maintained by the Biodiversity Information Standards (TDWG) formerly known as the Taxonomic Database Working Group which "focuses on the development of standards for the exchange of biological/biodiversity data" (TDWG, 2007). ABCD has been identified and recognized as a content standard which should be adopted and implemented for Diatomscapes and the remaining images of biological silica.

Curate and Preserve – Be aware of, and undertake management and administration actions planned to promote curation and preservation throughout the curation lifecycle.

✓**Diatomscapes –** After consultation with the FDA director, setup procedures instructions were sent on how to ingest, curate, and preserve Diatomscapes via FDA. DAITSS METS SIP Profile was created, FDA SIP Specification followed, and Ingest Reported generated as part of submitting materials to FDA for preservation. Real-time statistics are available online.

Sequential Actions

Conceptualize – Conceive and plan the creation of data, including capture method and storage options.

✓**Diatomscapes –** Diatomscapes were created by JSM-840 and FEI Nova Nano-400 electron scanning microscopes in the FSU Biological Science laboratory. Even though basic metadata has been captured, the need for a systematic approach to metadata creation and digital preservation were introduced to the researcher along with recommendation of using The DCC Curation Lifecycle Model as a framework reference tool to explore digital curation of images of biological silica.

Create or receive – Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation. Receive data, in accordance with documented policies, from data creators, other archives, other repositories or data centers, and if required assign appropriate metadata.

✓**Diatomscapes –** Some of the technical metadata created for Diatomscapes will be generated by a MD5 checksum, JSTOR/Harvard Object Validation Environment (JHOVE), and curation tools software applications.

Appraise & select – Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.

✓**Diatomscapes –** Diatomscapes images were selected and appraised by FSU biological scientists. Diatomscapes II images were selected based on the fact the images were created with FEI Nova Nano-400 electron scanning microscope which replaced and provide higher microscopic magnification than the legacy JSM-840 electron scanning microscope. Images were selected based on researchers' attitudes and experiences involved in the discovery and naming of some species.

Ingest – Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.

✓**Diatomscapes –** were FTP to the FCLA FDA server according to DAITSS METS SIP Profile and FDA SIP Specification as outlined in FDA Policy and Submitting Materials to the FDA (FDA, 2003).

Preservation action – Undertake actions to ensure long-term preservation and retention of authoritative nature of data. Preservation actions should ensure the data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.

✓**Diatomscapes –** MD5 checksum, JHOVE, and other curation tools will be used to create technical metadata which will then be included in the METS SIP Profile which describes the images of biological silica which are Diatomscapes.

Store – Store the data in a secure manner adhering to relevant standards.

✓**Diatomscapes –** The images of biological silica of Diatomscapes are store on a dark archive (non-web accessible) with technical and preservation metadata included in the DAITSS METS SIP which accompanied the digital objects on ingest.

Access, Use & Reuse – Ensure that data is accessible to both designated users, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.

✓**Diatomscapes –** Diatomscapes are in Flickr and Picasa with restrictive access (non-publicly viewable/accessible) until species name new to science are published. Some images of biological silica are also in Morphbank .

Transform – create new data from the original by migration into a different format or creating a subset by selection or query to create newly derived results, perhaps for publication (derivatives).

✓**Diatomscapes –** Diatomscapes tiff images were transformed into jpegs for ingest in Flickr and Picasa for sharing; into large-format posters for an exhibition display to promote faculty research; into a short mpeg video and uploaded to Facebook. Select images of biological silica were chosen to create Diatomscapes and Diatomscapes II. Some images of biological silica have been published in scholarly publications/journals.

Occasional Actions

Dispose – Dispose of data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian. In some instance data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction.

✓**Diatomscapes –** Disposition of images of biological silica from Diatomscapes has not yet been performed at the time of this paper.

Reappraise – Return data which fails validation procedures for further appraisal and selection.

✓**Diatomscapes –** Reappraise of images of biological silica from Diatomscapes are currently in review by scientists. Previous images of biological silica created with legacy scanning electron microscope may need to be rescanned with the current electron scanning for higher resolution and microscopic magnification.

Migrate – Migrate data to a different format. This may be done to accord with storage environment or to ensure the data's immunity from hardware or software obsolescence.

✓**Diatomscapes –** Diatomscapes images currently are preserved in the TIFF format. However, with the advancement in preservation standards comes emerging and new standards giving way to best practices. As a result, current preservation standards and file formats will be assessed bi-annually for compliance and to reduce the threat of technical obsolescence.

References:

Boyer, E. L. (1997). Scholarship reconsidered: priorities of the professoriate. Retrieved March 6, 2009 from http://www.pcrest.com/PC/FGB/test/2_5_1.htm.
DCC. (2008). The DCC curation lifecycle model. Retrieved December 9, 2008 from http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf.
Digital Curation Centre. (2008). Welcome. Retrieved December 9, 2008 from http://www.dcc.ac.uk/.
FDA. (2003). Florida digital archive. Retrieved December 10, 2008 from http://www.fcla.edu/digitalArchive/daInfo.htm.
Griffiths, A. (2008). The publication of research data: researcher attitudes and behavior. 4th international digital curation conference, Edinburgh, Scotland.
Higgins, S. (2007). Draft DCC curation lifecycle model. The International Journal of Digital Curation, Issue 2 (2), 2007. Retrieved December 9, 2008 from http://www.ijdc.net/ijdc/article/view/45/52.
Hunter, P. (2005, January). A tradition of scholarly documentation for digital objects: the launch of the digital curation centre. *Ariadne*, *42*. Retrieved September 18, 2007 from http://www.ariadne.ac.uk/issue42/dcc-rpt/.
Jantz, R. (2008). An institutional framework for creating authentic digital objects. 4th international digital curation conference, Edinburgh, Scotland.
JISC. (2003). JISC circular 6/03 (revised) digital curation centre. Retrieved December 9, 2008 from http://www.webarchive.org.uk/pan/13734/20060324/www.jisc.ac.uk/index2e1f.html?name=funding_digcentre
JISC. (2004). JISC e-Research brochure, e-Science data curation. Retrieved December 9, 2008 from www.dcc.ac.uk/docs/dcc-**life-cycle**.ppt.
Kintigh, K. (2006a). The promise and challenge of archeological data integration. *American Antiquity*, *71*(3), pp. 567-578.
Morphbank. (2008). About morphbank. Retrieved December 10, 2008 from http://www.morphbank.net/About/Introduction/.
RIN (2008) To Share or not to share: Publication and quality assurance of research data outputs. Research Information Network, June 2008. Retrieved December 10, 2008, from http://www.rin.ac.uk/files/Data%20publication%20report,%20main%20-%20final.pdf
TDWG. (2007). Biodiversity information standards (TWDG). Retrieved December 10, 2008 from http://www.tdwg.org/about-tdwg/.