

# Data Curation Workshop

## **Some Reflections on Students' Roles**

**ETD 2011: 14<sup>th</sup> Int. Symp. on ETDs  
Cape Town, South Africa**

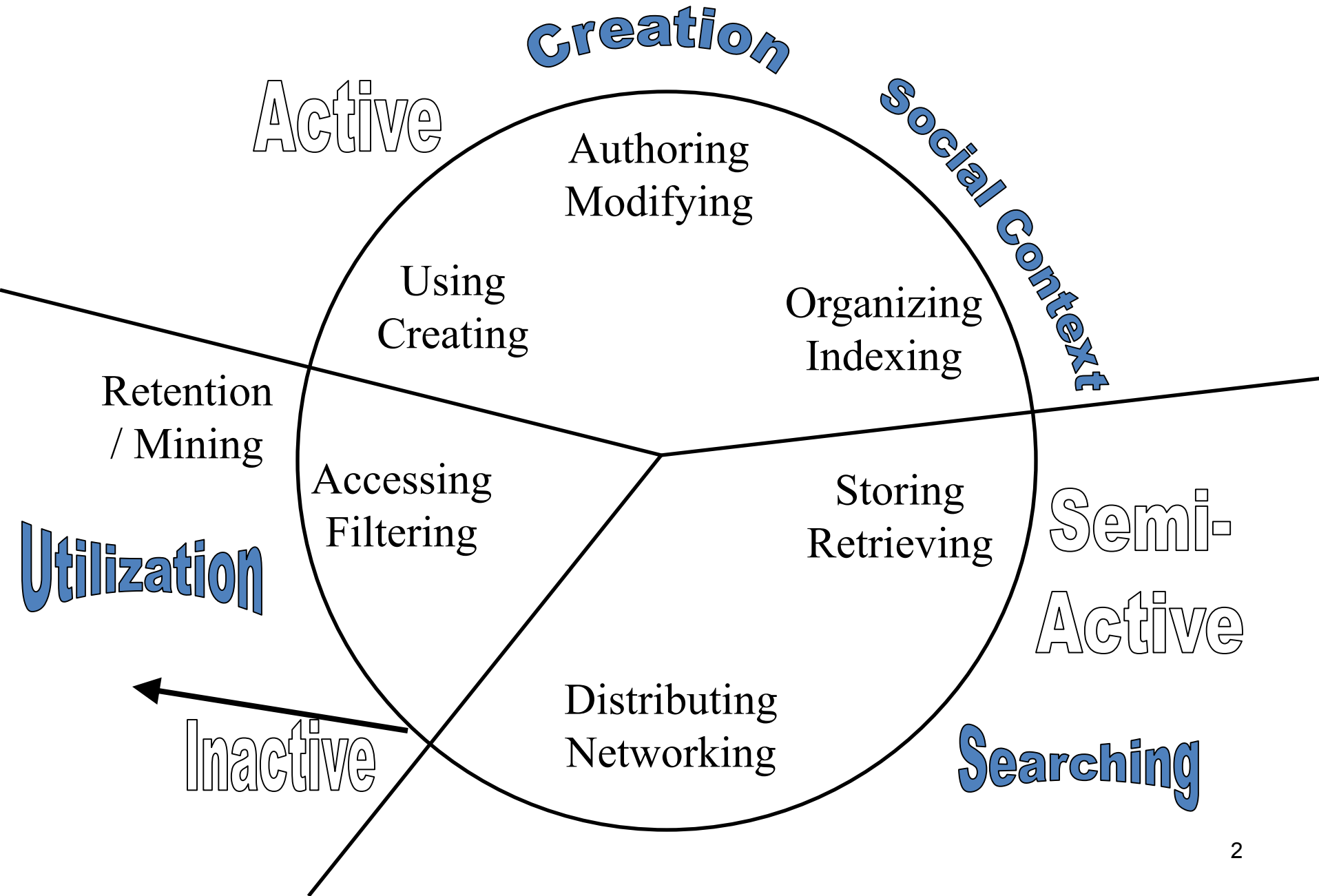
Edward A. Fox

Executive Director, NDLTD, [www.ndltd.org](http://www.ndltd.org)

[fox@vt.edu](mailto:fox@vt.edu)      <http://fox.cs.vt.edu/talks/2011>

Virginia Tech, Blacksburg, VA 24061 USA

# Information Life Cycle



# Case Study – Uma Murthy

- [umurthy21@gmail.com](mailto:umurthy21@gmail.com), [umurthy@vt.edu](mailto:umurthy@vt.edu)
- <http://scholar.lib.vt.edu/theses/available/etd-04142011-175752/>
- Fish image database
- Software (SuperIDR) w. data for SI, fish ID
- Experimental user study data
- Details of ETD (with multiple files)

# Uma's ETD Title (splash) page with initial portion of metadata, from April 2011

Firefox File Edit View History Bookmarks Tools Window Help

Title page for ETD etd-04142011-175752

NDLTD Board of Directors Meeti... x NDLTD Board of Directors — N... x CaringBridge / Alyssa Fox / Jou... x Fahrenheit to Celsius Conversio... x Current local time in South Afri... x Title page for ETD etd-041420... x ETD 2011 x +

http://scholar.lib.vt.edu/theses/available/etd-04142011-175752/

barbara evans ubc

**digital library and archives**  
*formerly the scholarly communications project*

**etds<sup>@vt</sup>**

**Title page for ETD etd-04142011-175752**

<b>Type of Document</b>	Dissertation												
<b>Author</b>	Murthy, Uma												
<b>Author's Email Address</b>	umurthy@vt.edu												
<b>URN</b>	etd-04142011-175752												
<b>Title</b>	Digital Libraries with Superimposed Information: Supporting Scholarly Tasks that Involve Fine Grain Information												
<b>Degree</b>	PhD												
<b>Department</b>	Computer Science												
<b>Advisory Committee</b>	<table><thead><tr><th>Advisor Name</th><th>Title</th></tr></thead><tbody><tr><td>Edward A. Fox</td><td>Committee Chair</td></tr><tr><td>Lois M. Delcambre</td><td>Committee Member</td></tr><tr><td>Manuel A. Perez-Quinones</td><td>Committee Member</td></tr><tr><td>Naren Ramakrishnan</td><td>Committee Member</td></tr><tr><td>Ricardo da Silva Torres</td><td>Committee Member</td></tr></tbody></table>	Advisor Name	Title	Edward A. Fox	Committee Chair	Lois M. Delcambre	Committee Member	Manuel A. Perez-Quinones	Committee Member	Naren Ramakrishnan	Committee Member	Ricardo da Silva Torres	Committee Member
Advisor Name	Title												
Edward A. Fox	Committee Chair												
Lois M. Delcambre	Committee Member												
Manuel A. Perez-Quinones	Committee Member												
Naren Ramakrishnan	Committee Member												
Ricardo da Silva Torres	Committee Member												
<b>Keywords</b>	<ul style="list-style-type: none"><li>• Annotation</li><li>• Digital libraries</li><li>• Fish species identification</li><li>• Image retrieval</li><li>• Metamodel</li><li>• Subdocument</li><li>• Superimposed information</li><li>• User study</li></ul>												

4

11

W P O

# Uma's ETD Title (splash) page with final portion of metadata, from April 2011

Firefox

File Edit View History Bookmarks Tools Window Help

Title page for ETD etd-04142011-175752

NDLTD Board of Directors Meeti... NDLTD Board of Directors — N... CaringBridge / Alyssa Fox / Jou... Fahrenheit to Celsius Conversio... Current local time in South Afri... Title page for ETD etd-041420... ETD 2011

http://scholar.lib.vt.edu/theses/available/etd-04142011-175752/ barbara evans ubc

User study

Date of Defense

2011-01-28

Availability

unrestricted


Abstract

Many scholarly tasks involve working with contextualized fine-grain information, such as a music professor creating a multimedia lecture on a musical style, while bringing together several snippets of compositions of that style. We refer to such contextualized parts of a larger unit of information (or whole documents), as subdocuments. Current approaches to work with subdocuments involve a mix of paper-based and digital techniques. With the increase in the volume and in the heterogeneity of information sources, the management, organization, access, retrieval, as well as reuse of subdocuments becomes challenging, leading to inefficient and ineffective task execution. A digital library (DL) facilitates management, access, retrieval, and use of collections of data and metadata through services. However, most DLs do not provide infrastructure or services to support working with subdocuments. Superimposed information (SI) refers to new information that is created to reference subdocuments in existing information resources. We combine this idea of SI with traditional DL services, to define and develop a DL with SI (an SI-DL). Our research questions are centered around one main question: how can we extend the notion of a DL to include SI, in order to support scholarly tasks that involve working with subdocuments? We pursued this question from a theoretical as well as a practical/user perspective. From a theoretical perspective, we developed a formal metamodel that precisely defines the components of an SI-DL, building upon related work in DLs, SI, annotations, and hypertext. From the practical/user perspective, we developed prototype superimposed applications and conducted user studies to explore the use of SI in scholarly tasks. We developed SuperIDR, a prototype SI-DL, which enables users to mark up subimages, annotate them, and retrieve information in multiple ways, including browsing, and text- and content-based image retrieval. We explored the use of subimages and evaluated the use of SuperIDR in fish species identification, a scholarly task that involves working with subimages. Findings from the user studies and other work in our research lead to theory- and experiment-based enhancements that can guide design of digital libraries with superimposed information.

Files

Filename	Size	Approximate Download Time (Hours:Minutes:Seconds)				
		28.8 Modem	56K Modem	ISDN (64 Kb)	ISDN (128 Kb)	Higher-speed Access
<a href="#">murthy-u-d-2011-dissertation.pdf</a>	26.22 Mb	02:01:22	01:02:25	00:54:37	00:27:18	00:02:19
<a href="#">murthy-u-d-2011-superidr-readme.pdf</a>	147.59 Kb	00:00:40	00:00:21	00:00:18	00:00:09	< 00:00:01
<a href="#">murthy-u-d-2011-superidr-source-images.zip</a>	77.91 Mb	06:00:41	03:05:29	02:42:18	01:21:09	00:06:55
<a href="#">murthy-u-d-2011-superidr-source-main.zip</a>	7.31 Mb	00:33:50	00:17:24	00:15:13	00:07:36	00:00:38

Browse All Available ETDs by ( [Author](#) | [Department](#) )

dlavirginia tech home

etds

image base

journals

contact dla

news

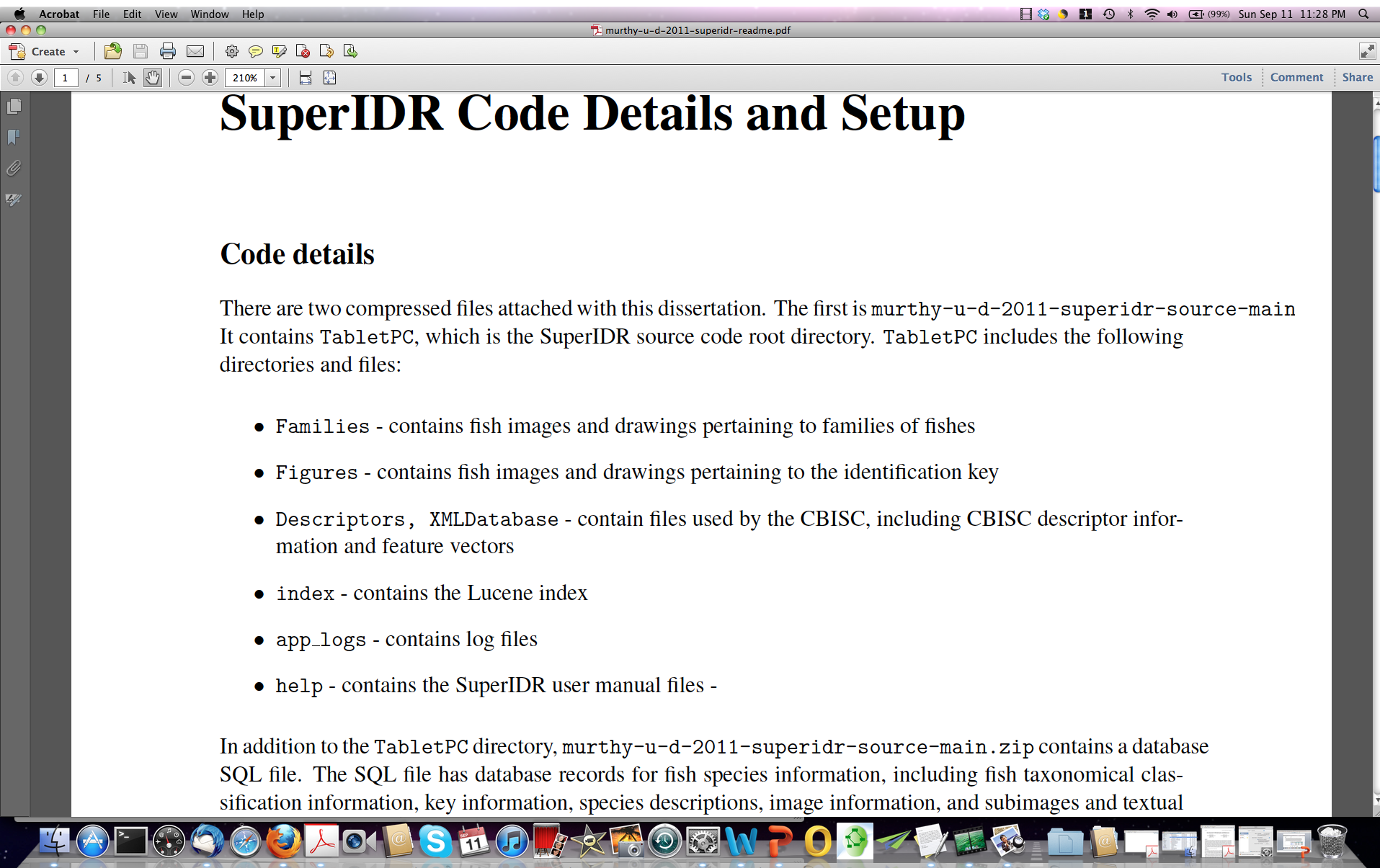
university libraries

online class materials

special collections

If you have questions or technical problems, please [Contact DLA](#).

# Uma's ETD Readme file, explaining the 2 compressed files, s/w, DB, from April 2011



## SuperIDR Code Details and Setup

### Code details

There are two compressed files attached with this dissertation. The first is `murthy-u-d-2011-superidr-source-main`. It contains TabletPC, which is the SuperIDR source code root directory. TabletPC includes the following directories and files:

- `Families` - contains fish images and drawings pertaining to families of fishes
- `Figures` - contains fish images and drawings pertaining to the identification key
- `Descriptors`, `XMLDatabase` - contain files used by the CBISC, including CBISC descriptor information and feature vectors
- `index` - contains the Lucene index
- `app_logs` - contains log files
- `help` - contains the SuperIDR user manual files -

In addition to the TabletPC directory, `murthy-u-d-2011-superidr-source-main.zip` contains a database SQL file. The SQL file has database records for fish species information, including fish taxonomical classification information, key information, species descriptions, image information, and subimages and textual

# CS Perspective

- Data
  - All types, for all types of research, application
  - Variety of standard, proprietary, new formats
  - Data, dataset, database, archived version
- Software
  - Dependencies on hardware, software
  - Frequent version shifts, making code obsolete
- Preservation: Raymond Lorie, IBM: UVC

# Needs, Problems

- NSF and others require a data management plan.
- Many research studies cannot be replicated since the student left, and with them went crucial information about data.
- There are no funds assigned to this work.
- Faculty lack time and knowledge.
- Few projects have professional staff to carry out this type of work.



# Data with ETDs

- Students are the naturals for this task.
- They can learn and contribute through working with content near and dear to them.
- They are the only ones in many cases who can provide the full provenance.
- It must be recorded before the student graduates, else this golden opportunity is lost.

# Conclusion

- We have a good case study – Uma Murthy.
- I argue that ETD authors are the natural ones to learn about and engage in data curation, as they finalize their ETDs.