3rd World Conference on AI, ML, DS

May 23, 2024

"A Digital Library to Promote Use of the World's Theses and Dissertations"

https://fox.cs.vt.edu/talks/2024/2024Eurasia3rdAIMLDSkeynoteFox.pdf

Keynote by Edward A. Fox, Ph.D., Professor

- fox@vt.edu http://fox.cs.vt.edu
- Dept of Computer Science (& ECE by courtesy)
- Virginia Tech, Blacksburg, VA 24061 USA
- NDLTD: Exec. Director, Chairman of the Board

Presentation Outline

- Acknowledgments
- NDLTD (Networked Digital Library of Theses and Dissertations: ndltd.org=theses.org)
- Intelligent Information Systems, 5S
- Digital Libraries
- Piloting: IMLS, CS5604 PBL instances
- Your To Do List

Acknowledgements (selected)

- Mentors (Licklider, Kessler, Salton); IMLS, NSF, and other sponsors
- Students, colleagues, co-investigators (selected): Lynn Abbott, Eman AbdelRahman, Aman Ahuja, Monika Akbar, Hamed Alhoori, Pranav Angara, Ashish Baghudana, Bipasha Banerjee, Elinor Benami, Warren Bickel, Boots Cassel, Saurabh Chakravarty, Prashant Chandrasekar, Satvik Chekuri, Yinlin Chen, Kiran Chitturi, Luis Escobar, Alexandre Falcao, Weiguo Fan, Francesco Ferretti, Eric Fouh, Chris Franck, Rick Furuta, Lee Giles, Marcos André Gonçalves, Doug Gorton, Islam Harb, S.M.Shamimul Hasan, Chongyu He, Michael Hsiao, Bill Ingram, Jeremy Jenrette, Palakh Jude, Adheesh Juvekar, Sampanna Kahu, Tarek Kanan, Ola Karajeh, Andrea Kavanaugh, Martin Klein, Harinni Kumar, Spencer Lee, Sunshin Lee, Jonathan Leidig, Lin Tzy Li, Liuqing Li, Yi Ma, Yufeng Ma, Mohamed Magdy, Shivam Maharshi, Madhav Marathe, Gary Marchionini, Paul Mather, Maanav Mehrotra, Uma Murthy, Pranav Nakate, Michael Nelson, Sanghee Oh, Sung Hee Park, Supritha Patil, Denilson Pereira, Jeff Pomerantz, Naren Ramakrishnan, Sagnik Ray-Choudhury, Chandan Reddy, Aditya Shah, Rao Shen, Ziqian Song, Venkat Srinivasan, Jeff Stein, Hussein Suleman, Allison Tegge, Ricardo Torres, Adithya Upadhya, Saket Vishwasrao, Xinyue Wang, Barbara Wildemuth, Jian Wu, Zhiwu Xie, Seungwon Yang, Xuan Zhang, ...

Related Funded Grants

1. IMLS LG-37-19-0078-19: Opening Books and the National Corpus of Graduate Research. 2019-2023. PI: William A. Ingram, Co-PIs: Edward A. Fox and Jian Wu: https://opening-etds.github.io/
2. Indo-US S&T Forum: Open Digital Libraries and Interoperability Workshop, 2003, PI Fox; Co-chairs: Shalini Urs, Mohammad Zubair, N. Balakrishnan
3. NSF IIS-0086227: Open Archives: Distributed services for physicists and graduate students (OAD): 2001-2004; PD Fox; German DFG PI E. Hilf
4. UNESCO: International Guide for the Creation of Electronic Theses and Dissertations: 12/28/2000-3/31/2002. PD (Project Director) E. Fox
5. SOLINET (Southeastern Library Network, USA): Networked Digital Library of Theses and Dissertations: 2000. Project director Fox
6. NSF IIS-0090153 (427963): US-Korea Joint Workshop on Digital Libraries: Removing Barriers to International Collaboration on Research and Education through Digital Libraries, 8/1/2000-9/30/2002. Project director Fox, co-PIs R.L. Larsen, R. W. Moore
7. U.S. Dept. of Education, FIPSE Program P116B61190: Improving Graduate Education with a National Digital Library of Theses and Dissertations: 1996-99; PIs Fox, J. Eaton, G. McMillan; support by SURA, Microsoft, Adobe

Selected ETD-related VT ETDs

- William Ingram, Digital Libraries of Electronic Theses and Dissertations, 2025
- Satvik Chekuri, Scholarly Information System for Long Documents and their Elements, 2025
- Bipasha Banerjee. Improving access to ETD elements through chapter categorization and summarization, Summer 2024
- Aman Ahuja. Analyzing and Navigating Electronic Theses and Dissertations", 2023
- Dhanush Dinesh, Utilizing Docker and Kafka for Highly Scalable Bulk Processing of Electronic Theses and Dissertations (ETDs), 2023
- Javaid Manzoor, Segmenting Electronic Theses and Dissertations by Chapters, 2022
- Prashant Chandrasekar, Continuously Extensible Information Systems: Extending the 5S Framework by Integrating UX and Workflows, 2021
- Sampanna Kahu, Figure extraction from scanned electronic theses and dissertations,2020
- Palakh Mignonne Jude, Increasing Accessibility of Electronic Theses and Dissertations (ETDs) Through Chapter-level Classification, 2020
- Sung Hee Park, Discipline-Independent Text Information Extraction from Heterogeneous Styled References Using Knowledge from the Web, 2013
- W. Ryan Richardson, Using Concept Maps as a Tool for Cross-Language Relevance Determination, 2007
- Douglas Gorton, Practical Digital Library Generation into DSpace with the 5S Framework, 2007
- Hussein Suleman, Open Digital Libraries, 2002

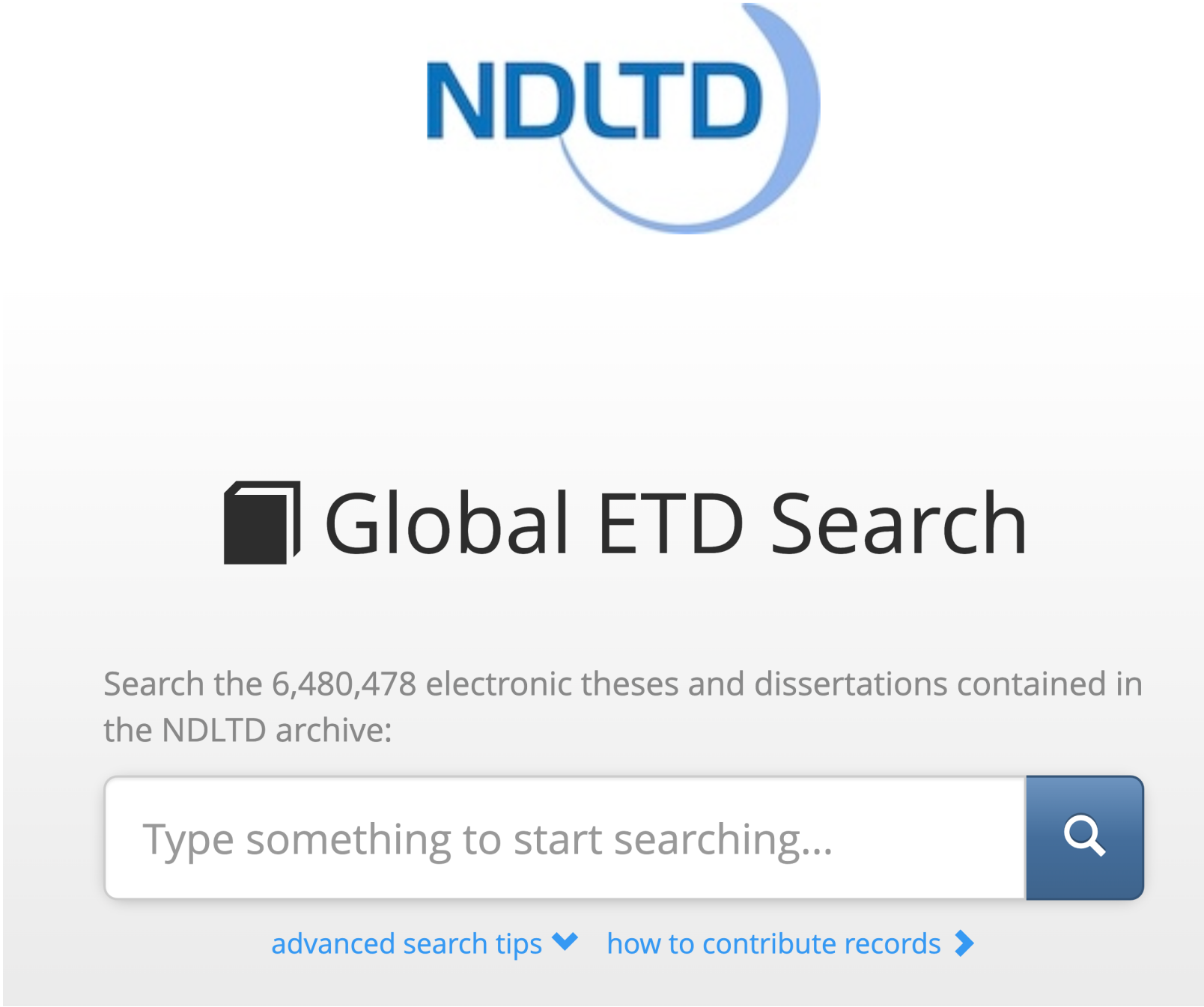
NDLTD: Mission

The Networked Digital Library of Theses and Dissertations (NDLTD) is an international organization dedicated to promoting the adoption, creation, use, dissemination, and preservation of electronic theses and dissertations (ETDs). We support electronic publishing and open access to scholarship in order to enhance the sharing of knowledge worldwide. Our website includes resources for university administrators, librarians, faculty, students, and the general public. Topics include how to find, create, and preserve ETDs; how to set up an ETD program; legal and **technical questions**; and the latest news and research in the ETD community.

J-ETD.org, j-etd@ndltd.org
Journal of Electronic Theses and Dissertations

- Open-access launch 1/1/2021. Please support!
- Managing Editor: Charles J. Greenberg
- Executive Editor: Edward A. Fox; Associate Editors : Suzanne Lorraine (Suzie) Allard (USA), Ramesh C. Gaur (India), Charles J. Greenberg (USA), Libio Huaroto (Peru), William A. Ingram (USA), Ana Sofia de Sousa Machado Mota (Portugal), Prashant Pandey (Australia), Ana Pavani (Brazil), Joachim Schöpfel (France), Janette Wright (UAE)

search.ndltd.org



Scenarios of Future Use of ETD DLs

1. Open problem -> plan for research
2. Problem -> list of references, related ETDs
3. Bibliography -> clusters -> lit. review chapter
4. Course (e.g., seminar) units based on ETDs
5. Final defense -> told missing cites of related ETDs
6. Promotion: impact of candidate's students' ETDs
7. Research trends: classification, topic modeling
8. Analysis & Assessment -> logs -> use by:
   - Local grad students, faculty, undergrads
   - Graduate School, Registrar, Research Division

Scenarios of Future Use:

Example: Open problem -> plan for research

1. Student volunteers to pilot test the new DL
2. Goal: find problem to solve
3. Explains her interest and background
4. Receives extracts from related ETDs:
   - open problems, planned future work
5. Selects top 5
6. Receives related ETD list, with chapter summaries
7. Fetches and studies top 2 ETDs from the list
8. Meets advisor to devise research plan

Intelligent Information Systems

- Digital libraries, repositories
- Search engines, recommenders
- Chatbots
- Bots
- Smart homes, smart cities
- Robots
- (Semi)autonomous vehicles, UAVs
  - Land: Cars, Trucks
  - Air: Planes, Drones
  - Sea: Boats, UUVs

5S Layers

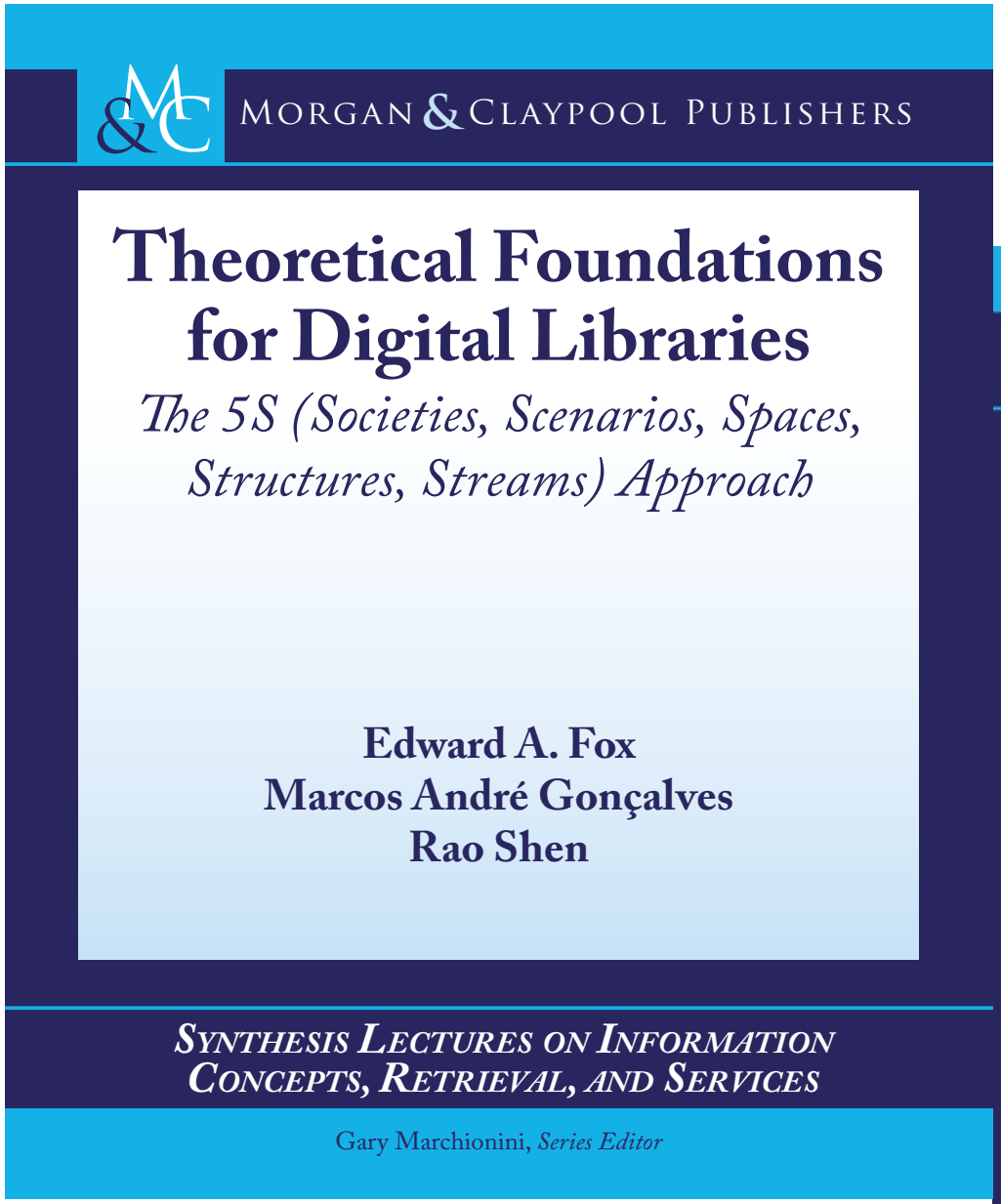
Societies

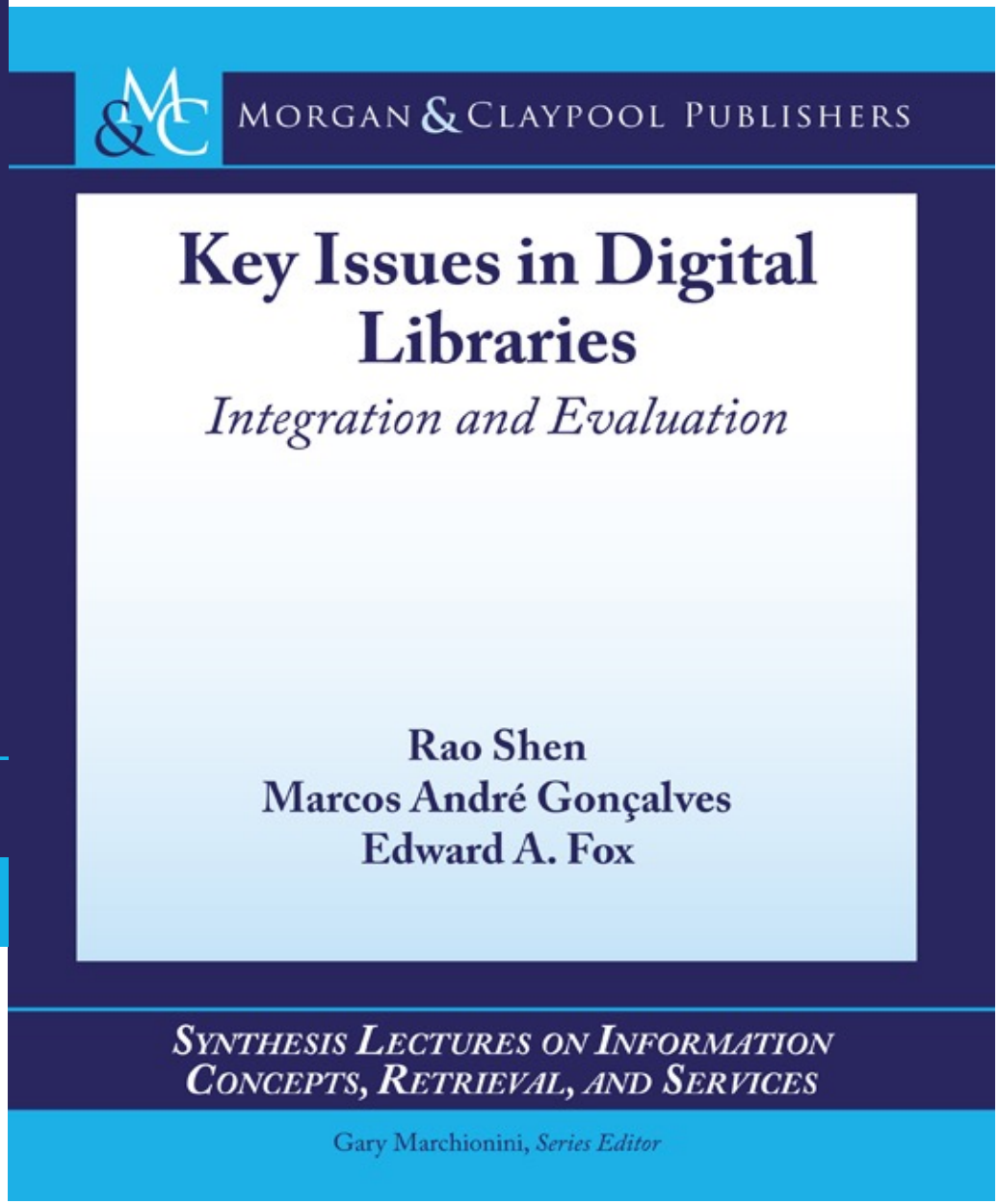
Scenarios

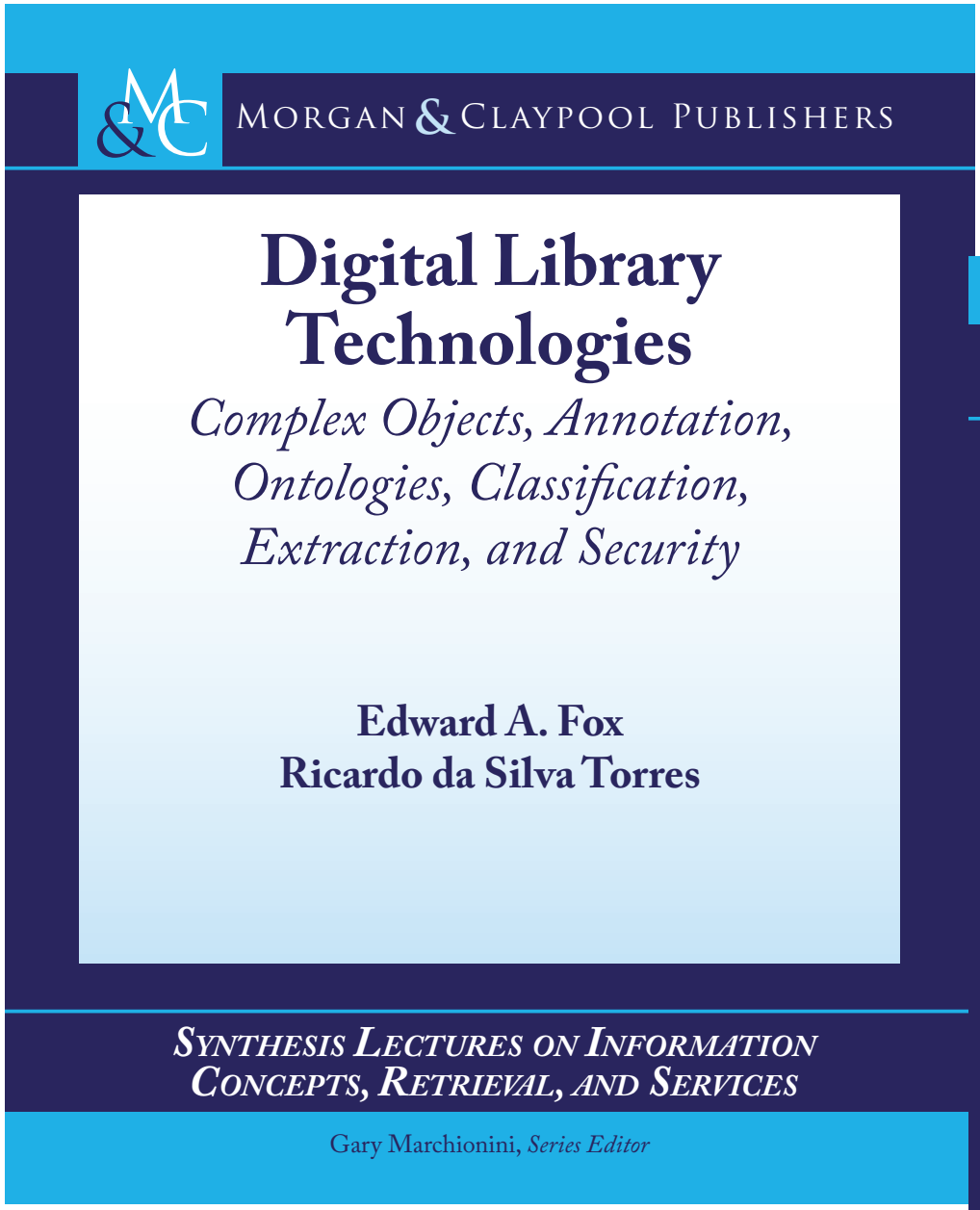
Spaces

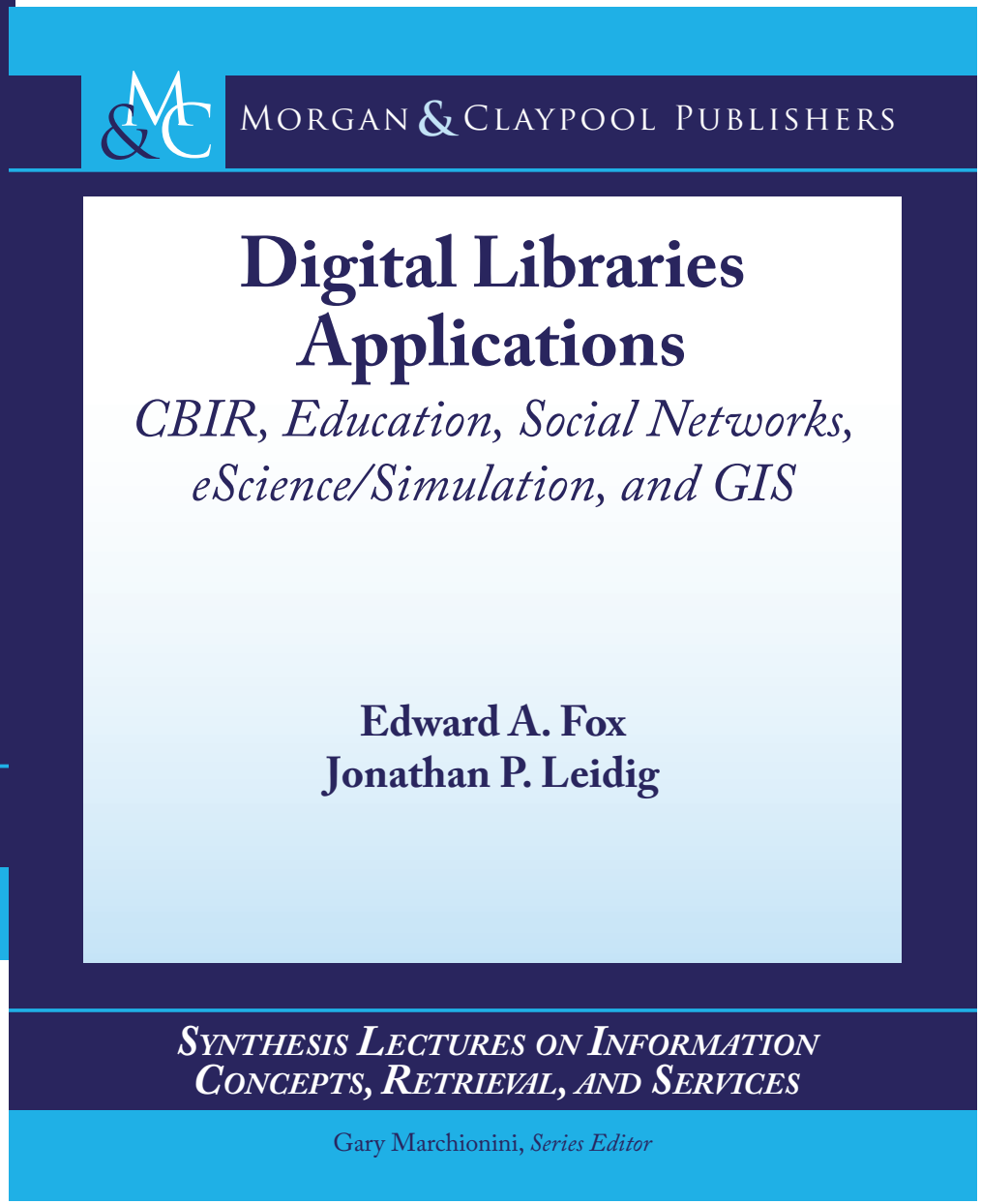
Structures

Streams


MORGAN & CLAYPOOL PUBLISHERS
Theoretical Foundations for Digital Libraries
The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach
Edward A. Fox
Marcos André Gonçalves
Rao Shen
SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES
Gary Marchionini, Series Editor


MORGAN & CLAYPOOL PUBLISHERS
Key Issues in Digital Libraries
Integration and Evaluation
Rao Shen
Marcos André Gonçalves
Edward A. Fox
SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES
Gary Marchionini, Series Editor


MORGAN & CLAYPOOL PUBLISHERS
Digital Library Technologies
Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security
Edward A. Fox
Ricardo da Silva Torres
SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES
Gary Marchionini, Series Editor

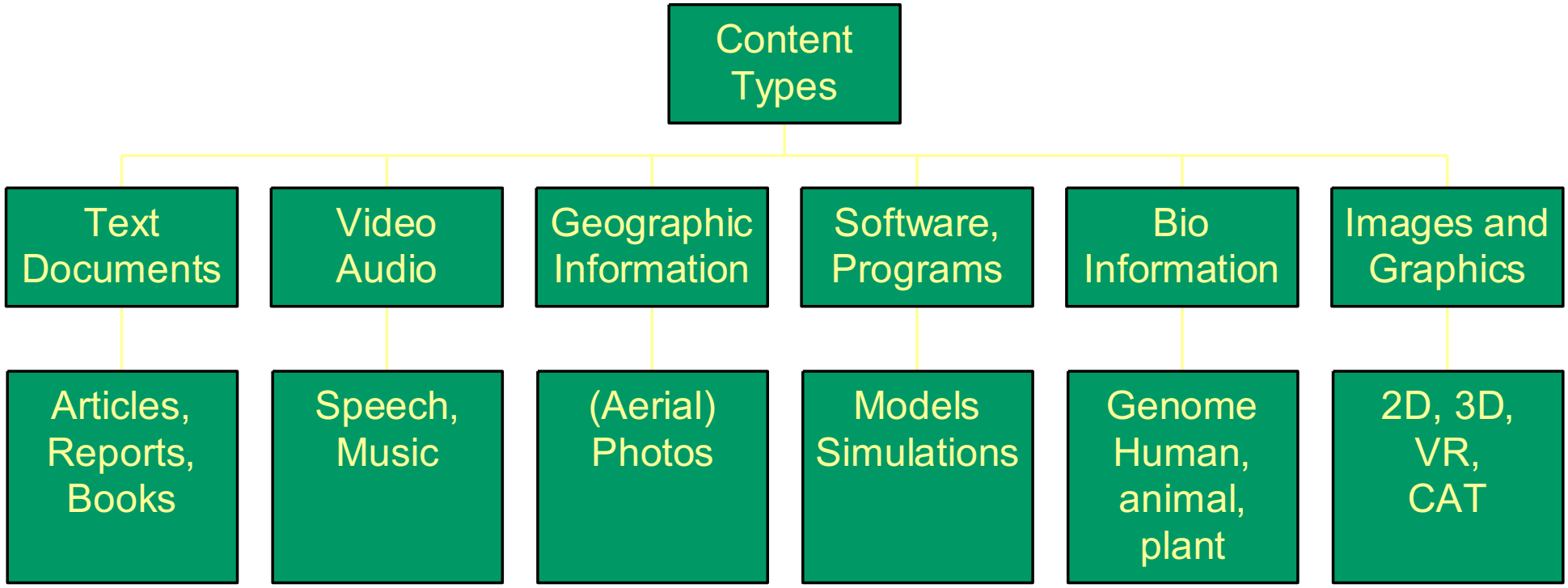

MORGAN & CLAYPOOL PUBLISHERS
Digital Libraries Applications
CBIR, Education, Social Networks, eScience/Simulation, and GIS
Edward A. Fox
Jonathan P. Leidig
SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES
Gary Marchionini, Series Editor

Informal 5S & DL Definitions

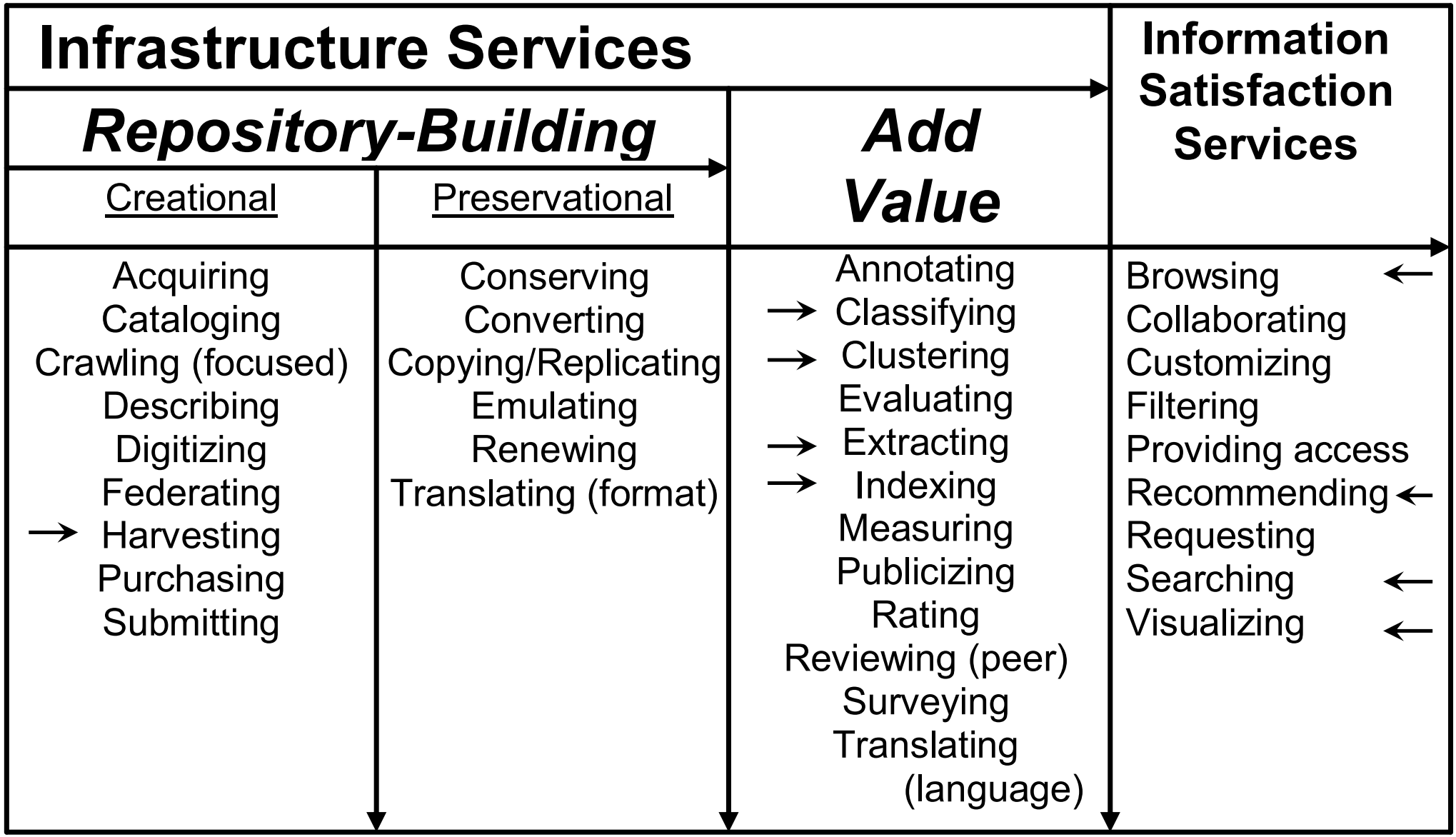
DLs are complex systems that

- help satisfy info needs of users (**societies**)
- provide info services (**scenarios**)
- organize info in usable ways (**structures**)
- present info in usable ways (**spaces**)
- communicate info with users (**streams**)

Digital Libraries: Content



Supporting Services across the Lifecycle



Quality Dimensions

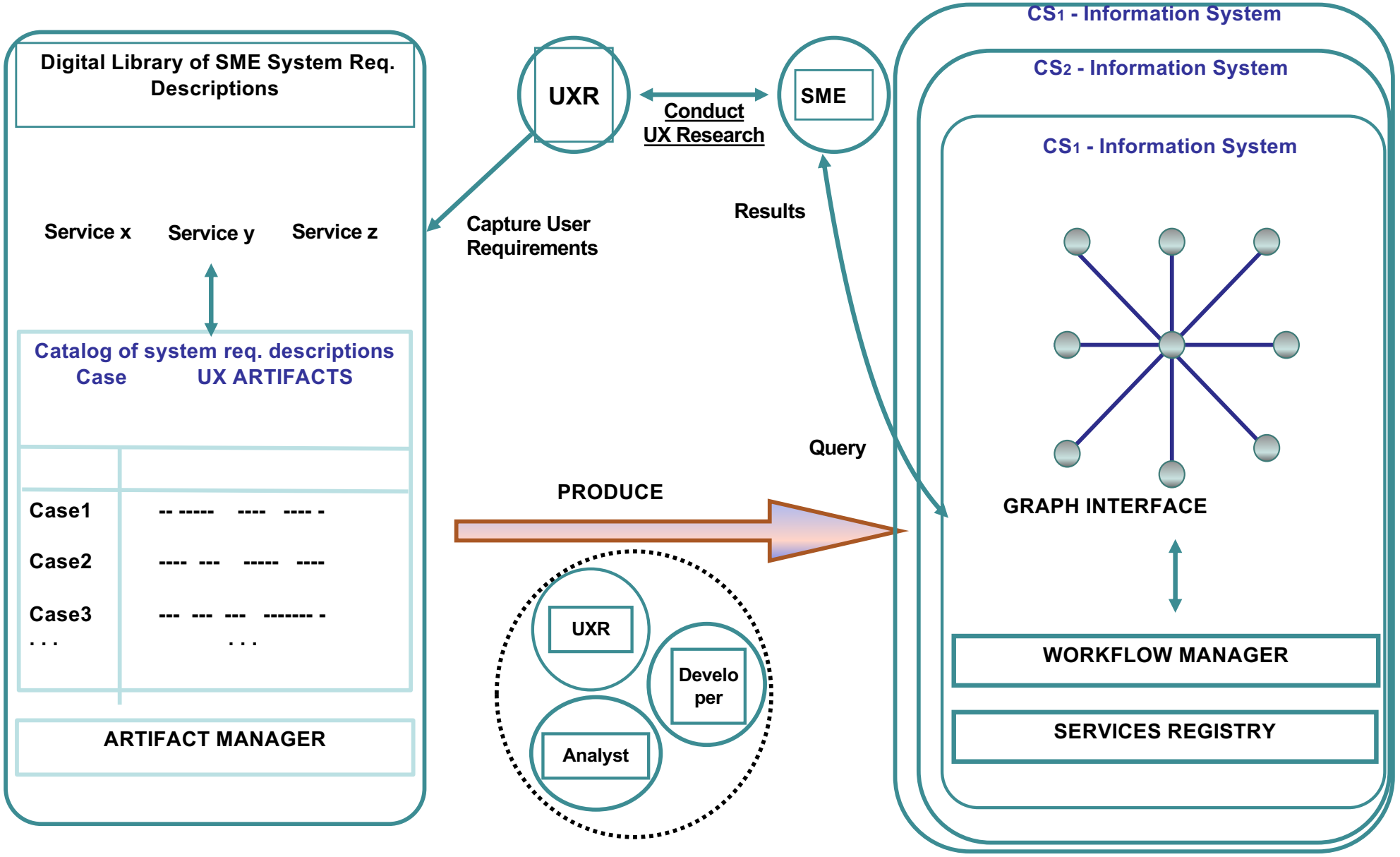
DL Concept	Dimensions of Quality
Digital object	Accessibility Pertinence Preservability Relevance Similarity Significance Timeliness
Metadata specification	Accuracy Completeness Conformance
Collection	Completeness Impact Factor
Catalog	Completeness Consistency
Repository	Completeness Consistency
Services	Composability Efficiency Effectiveness Extensibility Reusability Reliability

Scenarios of Future Use / Building DLs

1. UX: Customer discovery: subject-matter experts
2. UX: Validated list of:
   - Jobs-to-be-done, tasks, sub-tasks, goals, sub-goals
3. Personas
   1. Curators
   2. Experimenters
   3. Researchers (students, faculty, …)
4. DL software developer: knowledge graph mapping:
   - Goals, Sub-goals, Tasks, Sub-tasks
   - Workflows of services: Existing, Desired
5. Operations (Docker, Airflow; DevOps with CI/CD)

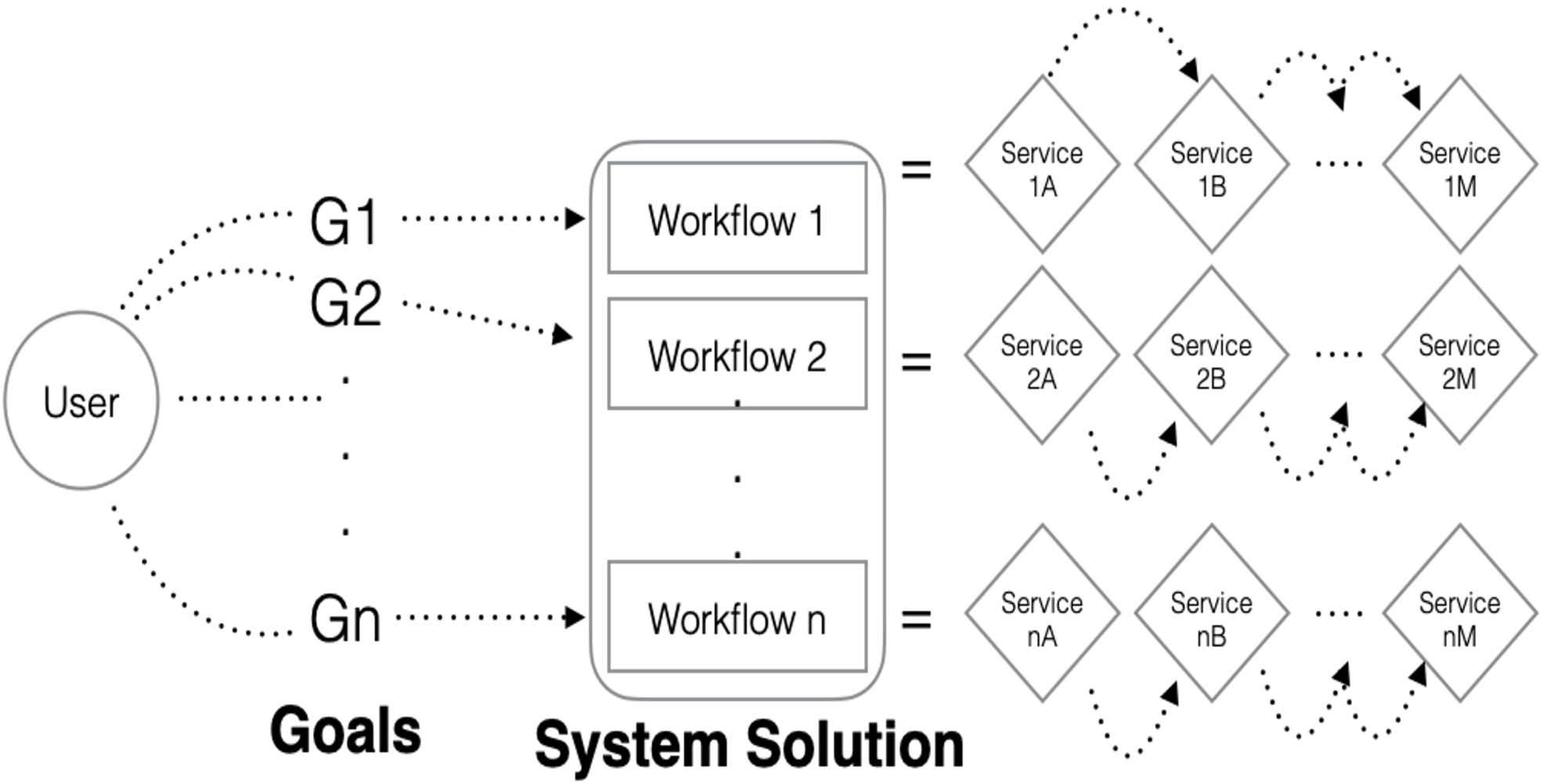
(Doctoral work of Prashant Chandrasekar)

Prashant Chandrasekar's DL Architecture



http://hdl.handle.net/10919/103815, https://doi.org/10.2478/dim-2021-0003

Workflow-defining Goal Decomposition



Opening Graduate Research

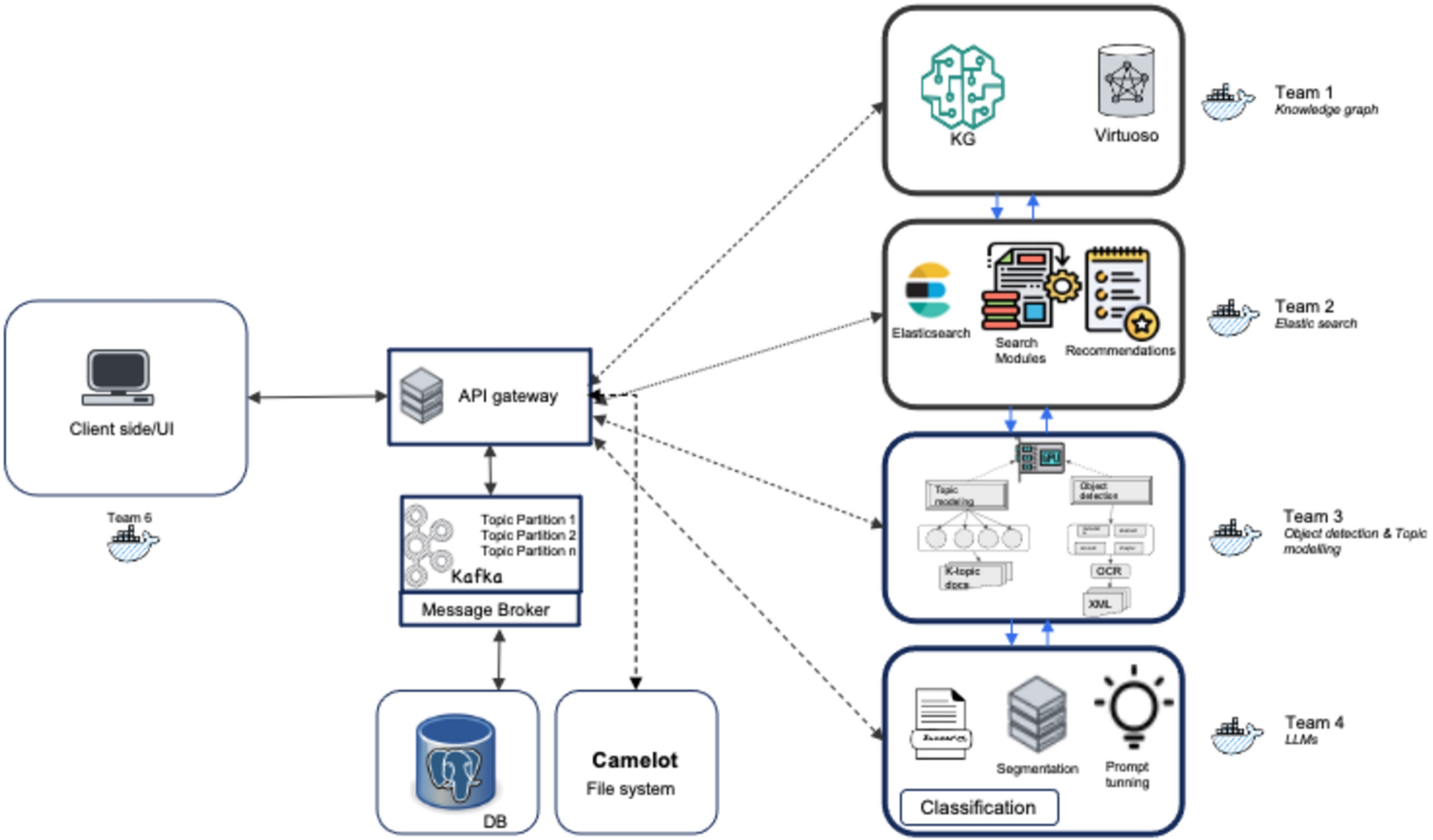
IMLS; 2019-2023; PI: William Ingram

- Activities
  - Collecting: 500,000+ from USA
    - Large universities, HBCUs, HSIs + Arabic corpus
  - Analyzing: parsing / detecting (texts, images)
  - Extracting: tables, figures, equations, references…
  - Scanned ETDs -> improved metadata
  - Classification, Topic Modeling -> Browsing
  - Segmenting: chapters -> Chapter summaries
- Results: New methods & technologies, pilot system (search, browse, recommend, viz)

CS5604 SMEs

- Aman Ahuja: topic modeling, object detection/document parsing (http://hdl.handle.net/10919/115817)
- Bipasha Banerjee, Sara Ahmadi: segmentation, language models, transformers, classification, summarization
- Prashant Chadrasekar, Dhanush Dinesh: integration, workflows, extensibility, DevOps
- Satvik Chekuri: search, recommendation
- Sung Hee Park, Bill Ingram: database, files

CS5604 Fall 2023 Teams



CS5604 ETD Team: Figure Extraction

Inference is accomplished via the best performing model trained by Sampanna and others

microscope observations of live bundles, and studies of kinocilium height (Fontilla and Peterson, 2000), were used to define heights of stereocilia and the kinocilium. The height data was obtained from various bundles that were different from, but similar to, the original bundle. In this manner a realistic representation of a bundle was assembled. The computer-generated graphic for each bundle in Figure 2.2 is based on the model input into *bmod*, and shows the deformed state of the bundle. Although it may not be clear from Figure 2.2, cells 1, 2, 4, and 5 are "loose-packed", and cells 3 and 6 are "tight-packed", as defined in Chapter 1.

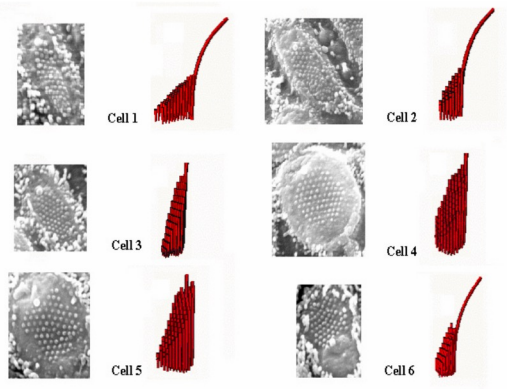


Figure 2.2: Six utricle cells – electron micrograph and 3-D rendering

Obviously, many approximations were made in modeling the cell bundles. Stereocilia diameters and spacing were approximated as constant throughout a given bundle. Perfect hexagonal layouts do not exist in biological bundles, but they are much easier to model. Cilia heights were based on similar bundles, and were approximated so as to linearly decrease in height along the E-I axis. Tapering at the base of stereocilia was

Page image(s)

microscope observations of live bundles, and studies of kinocilium height (Fontilla and Peterson, 2000), were used to define heights of stereocilia and the kinocilium. The height data was obtained from various bundles that were different from, but similar to, the original bundle. In this manner a realistic representation of a bundle was assembled. The computer-generated graphic for each bundle in Figure 2.2 is based on the model input into *bmod*, and shows the deformed state of the bundle. Although it may not be clear from Figure 2.2, cells 1, 2, 4, and 5 are "loose-packed", and cells 3 and 6 are "tight-packed", as defined in Chapter 1.

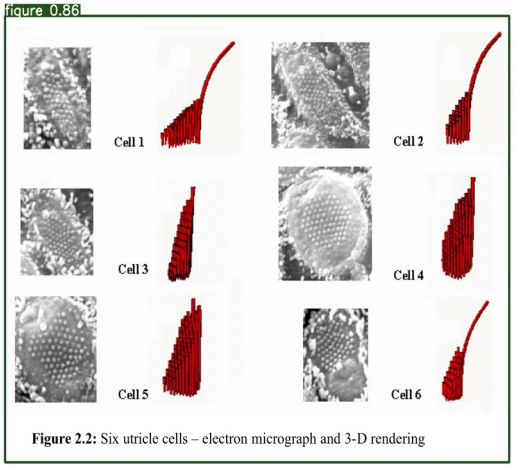


Figure 2.2: Six utricle cells – electron micrograph and 3-D rendering

Obviously, many approximations were made in modeling the cell bundles. Stereocilia diameters and spacing were approximated as constant throughout a given bundle. Perfect hexagonal layouts do not exist in biological bundles, but they are much easier to model. Cilia heights were based on similar bundles, and were approximated so as to linearly decrease in height along the E-I axis. Tapering at the base of stereocilia was

CHAPTER 2: METHODS AND MATERIALS 18

Page image(s) with image bound information

CS5604 ETD Team: Cropping

microscope observations of live bundles, and studies of kinocilium height (Fontilla and Peterson, 2000), were used to define heights of stereocilia and the kinocilium. The height data was obtained from various bundles that were different from, but similar to, the original bundle. In this manner a realistic representation of a bundle was assembled. The computer-generated graphic for each bundle in Figure 2.2 is based on the model input into *bmod*, and shows the deformed state of the bundle. Although it may not be clear from Figure 2.2, cells 1, 2, 4, and 5 are "loose-packed", and cells 3 and 6 are "tight-packed", as defined in Chapter 1.

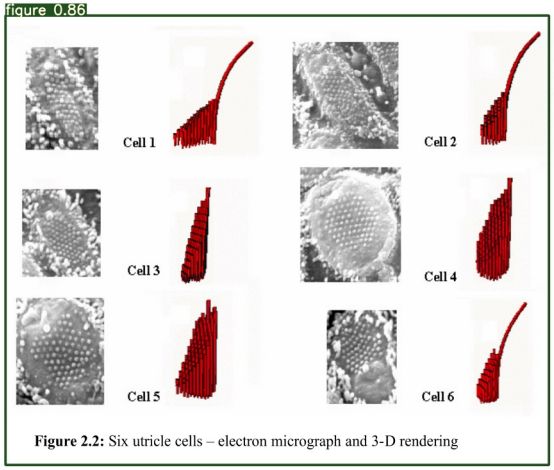


Figure 2.2: Six utricle cells – electron micrograph and 3-D rendering

Obviously, many approximations were made in modeling the cell bundles. Stereocilia diameters and spacing were approximated as constant throughout a given bundle. Perfect hexagonal layouts do not exist in biological bundles, but they are much easier to model. Cilia heights were based on similar bundles, and were approximated so as to linearly decrease in height along the E-I axis. Tapering at the base of stereocilia was

Page image(s) with image bound information

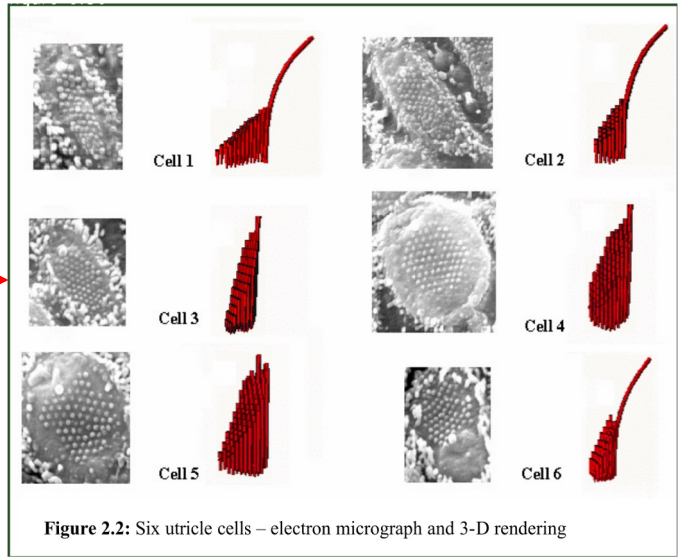


Figure 2.2: Six utricle cells – electron micrograph and 3-D rendering

Cropped images

CS5604 ETD Team: Ch. Segmentation

Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

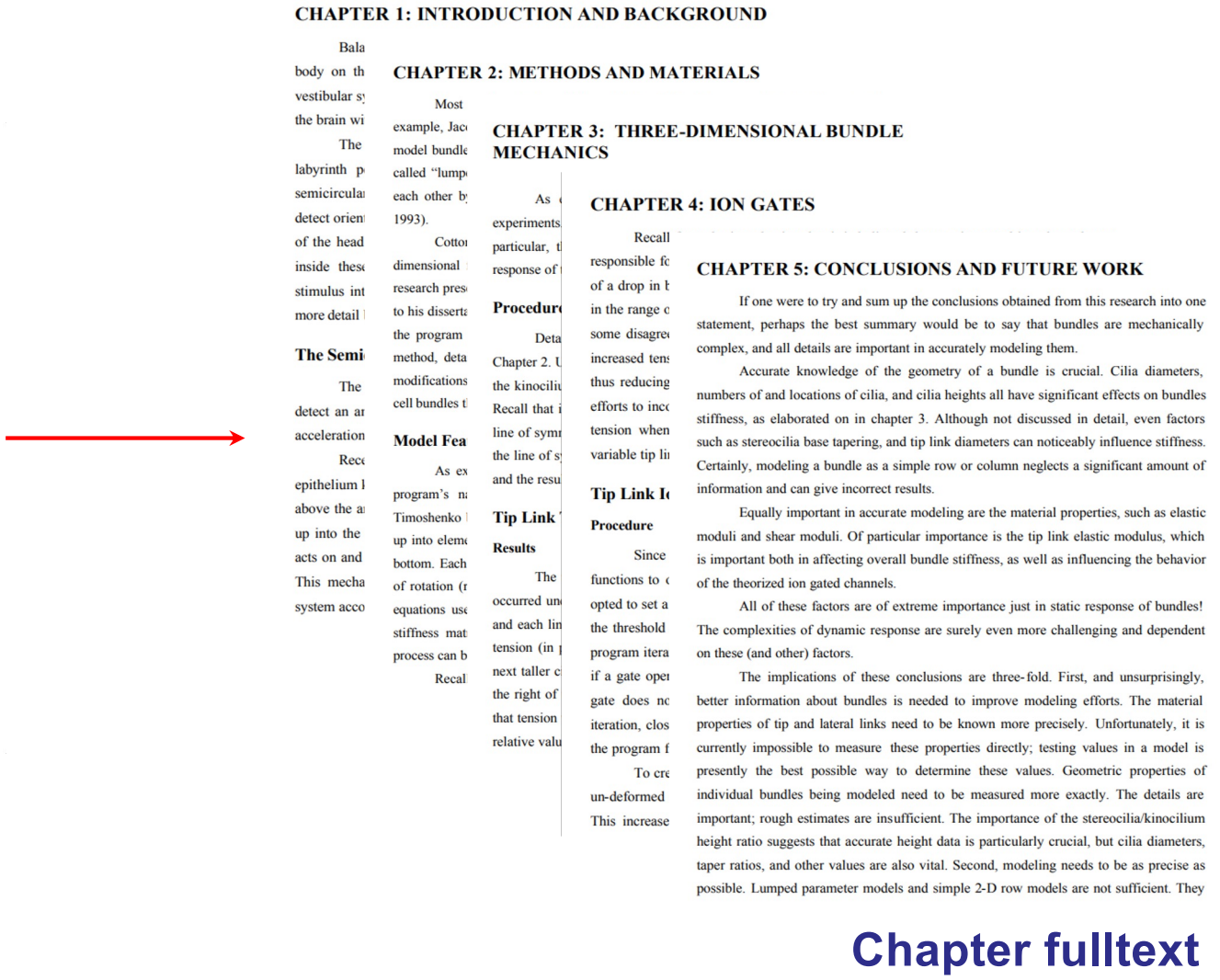
in

Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element



CHAPTER 1: INTRODUCTION AND BACKGROUND

CHAPTER 2: METHODS AND MATERIALS

CHAPTER 3: THREE-DIMENSIONAL BUNDLE MECHANICS

CHAPTER 4: ION GATES

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

If one were to try and sum up the conclusions obtained from this research into one statement, perhaps the best summary would be to say that bundles are mechanically complex, and all details are important in accurately modeling them.

Accurate knowledge of the geometry of a bundle is crucial. Cilia diameters, numbers of and locations of cilia, and cilia heights all have significant effects on bundles stiffness, as elaborated on in chapter 3. Although not discussed in detail, even factors such as stereocilia base tapering, and tip link diameters can noticeably influence stiffness. Certainly, modeling a bundle as a simple row or column neglects a significant amount of information and can give incorrect results.

Equally important in accurate modeling are the material properties, such as elastic moduli and shear moduli. Of particular importance is the tip link elastic modulus, which is important both in affecting overall bundle stiffness, as well as influencing the behavior of the theorized ion gated channels.

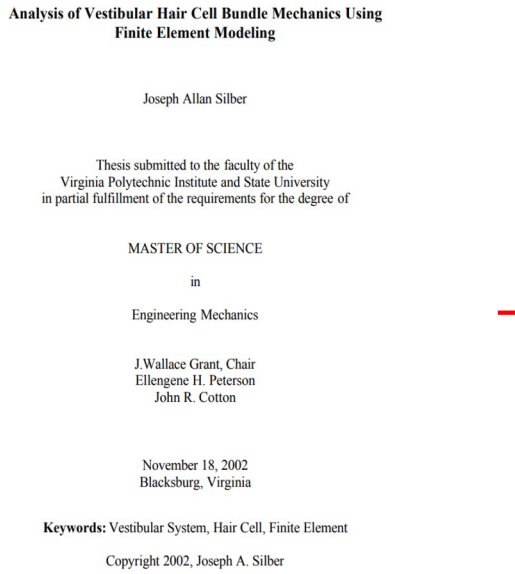
All of these factors are of extreme importance just in static response of bundles! The complexities of dynamic response are surely even more challenging and dependent on these (and other) factors.

The implications of these conclusions are three-fold. First, and unsurprisingly, better information about bundles is needed to improve modeling efforts. The material properties of tip and lateral links need to be known more precisely. Unfortunately, it is currently impossible to measure these properties directly; testing values in a model is presently the best possible way to determine these values. Geometric properties of individual bundles being modeled need to be measured more exactly. The details are important; rough estimates are insufficient. The importance of the stereocilia/kinocilium height ratio suggests that accurate height data is particularly crucial, but cilia diameters, taper ratios, and other values are also vital. Second, modeling needs to be as precise as possible. Lumped parameter models and simple 2-D row models are not sufficient. They

Chapter fulltext

CS5604 ETD Team: Text Extraction

Extract text from PDF page. Write the text into a .txt file.
Repeat until finish all PDF pages.



Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

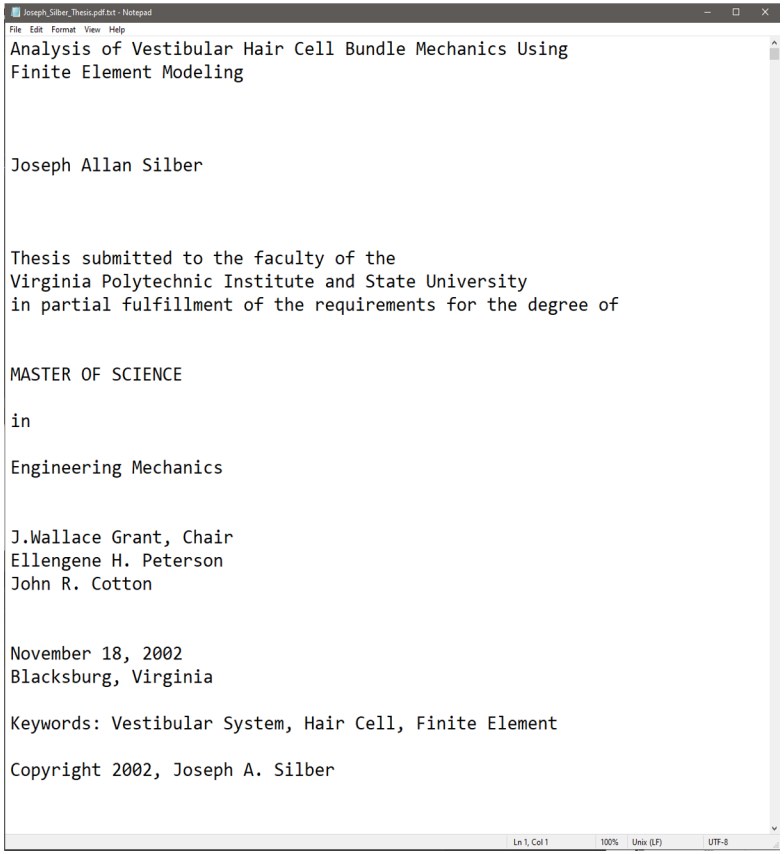
Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

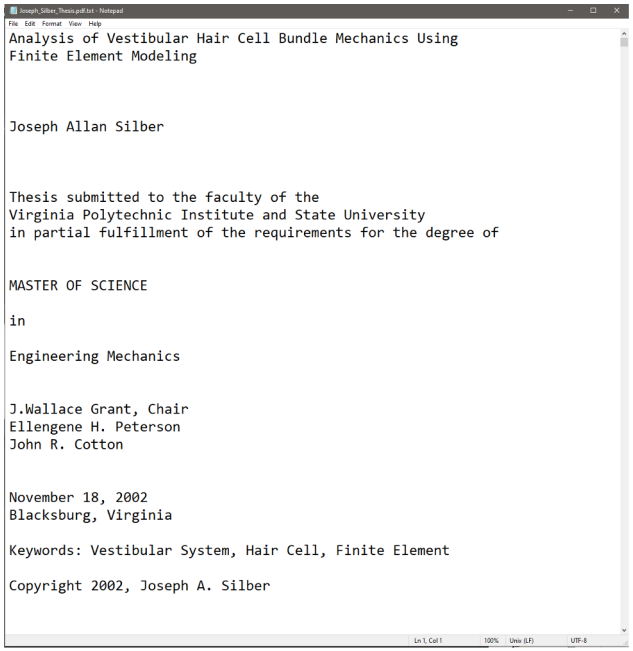
Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silber



Extracted Text

CS5604 ETD Team: Classification



Analysis of Vestibular Hair Cell Bundle Mechanics Using
Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Engineering Mechanics

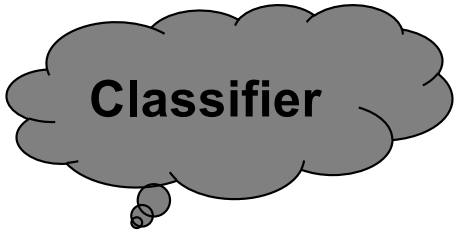
J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

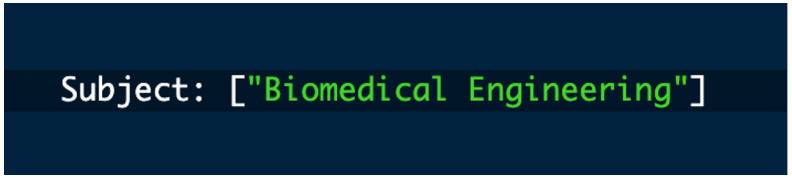
November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

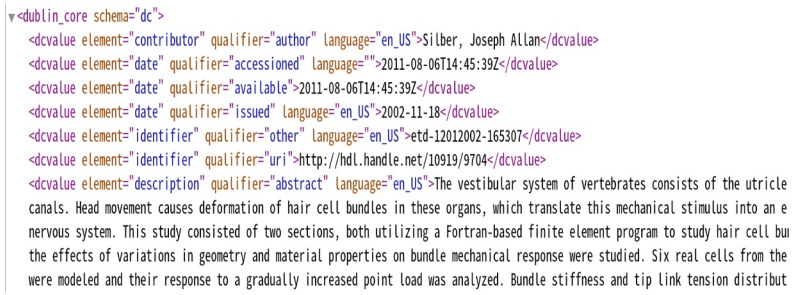
Copyright 2002, Joseph A. Silber

Extracted Text





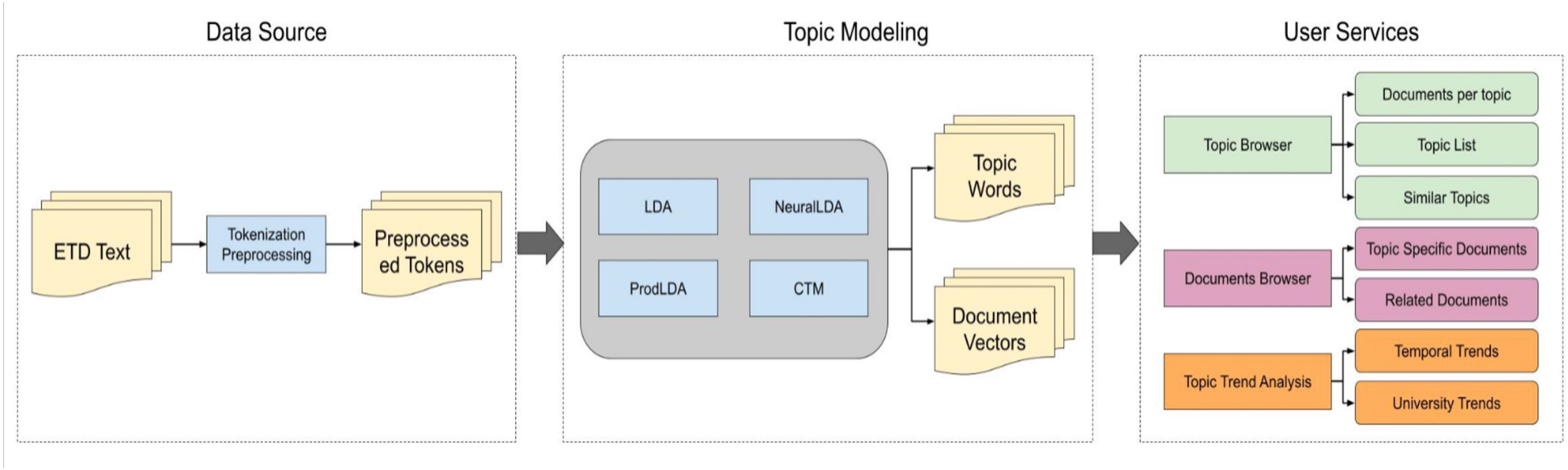
Labels for ETD



```
<dublin_core schema="dc">
  <dcvalue element="contributor" qualifier="author" language="en_US">Silber, Joseph Allan</dcvalue>
  <dcvalue element="date" qualifier="accessioned" language="">2011-08-06T14:45:39Z</dcvalue>
  <dcvalue element="date" qualifier="available">2011-08-06T14:45:39Z</dcvalue>
  <dcvalue element="date" qualifier="issued" language="en_US">2002-11-18</dcvalue>
  <dcvalue element="identifier" qualifier="other" language="en_US">etd-12012002-165307</dcvalue>
  <dcvalue element="identifier" qualifier="uri">http://hdl.handle.net/10919/9704</dcvalue>
  <dcvalue element="description" qualifier="abstract" language="en_US">The vestibular system of vertebrates consists of the utricle
  canals. Head movement causes deformation of hair cell bundles in these organs, which translate this mechanical stimulus into an e
  nervous system. This study consisted of two sections, both utilizing a Fortran-based finite element program to study hair cell bu
  the effects of variations in geometry and material properties on bundle mechanical response were studied. Six real cells from the
  were modeled and their response to a gradually increased point load was analyzed. Bundle stiffness and tip link tension distribut
```

Dublin Core XML

Example: ETD-Topics (Architecture)



Aman Ahuja, William A. Ingram, Chenyu Mao, Chongyu He, Jianchi Wei and Edward A. Fox.
Analyzing and Navigating ETDs Using Topic Models.
ETD 2022 conference, Novi Sad, Serbia, September 7-9, 2022

Your To Do List

- Use, and Encourage Others to Use:

  http://search.ndltd.org/
- Support NDLTD(.org) and Worldwide ETD Requirements
- Use Digital Libraries and Repositories
- Improve Information Systems using 5S & AI
- Apply PBL in Courses (or Help as a SME)

Questions?
Discussion?

Thank You!

fox@vt.edu

fox@ndltd.org

j-etd@ndltd.org

https://fox.cs.vt.edu/cv.htm