Introduction to Digital Libraries

Edward A. Fox
Department of Computer Science
Virginia Tech
Blacksburg, VA, 24061 USA
fox@vt.edu, ORCID 0000-0003-1447-6870

Yinlin Chen
University Libraries
Virginia Tech
Blacksburg, VA, 24061 USA
ylchen@vt.edu

***Abstract*—This tutorial is a thorough and deep introduction to the Digital Libraries (DL) field, providing a firm foundation: covering key concepts and terminology, as well as services, systems, technologies, methods, standards, projects, issues, and practices. It introduces and builds upon a firm theoretical foundation (starting with the '5S' set of intuitive aspects: Streams, Structures, Spaces, Scenarios, Societies), giving careful definitions and explanations of all the key parts of a 'minimal digital library', and expanding from that basis to cover key DL issues. Illustrations come from a set of case studies, including from multiple current projects, including with the application of natural language processing and machine learning to webpages, tweets, and long documents. Attendees will be exposed to four Morgan and Claypool books that elaborate on 5S. Further, new material will be added on building digital libraries using container and cloud services, on developing a digital library for electronic theses and dissertations, and methods to integrate UX and DL design approaches.**

I. Instructors

Professor Edward A. Fox holds a Ph.D. and M.S. in Computer Science from Cornell University, and a B.S. from M.I.T. He is a Fellow of both ACM and IEEE, cited for leadership in digital libraries and information retrieval, and a member of the SIGIR Academy. He serves as Executive Director and Chairman of the Board of the Networked Digital Library of Theses and Dissertations.

Since 1983 he has been at Virginia Polytechnic Institute and State University (VPI&SU or Virginia Tech), where he serves as Professor of Computer Science, and by courtesy, of ECE. He directs the Digital Library Research Laboratory. He was an elected member of the Board of Directors of the Computing Research Association and was Chair (now a member) of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) Steering Committee as well as of the IEEE-CS Technical Committee on Digital Libraries. With support from 138 research grants, he has worked in the areas of digital libraries, information storage and retrieval, machine learning (ML) /AI, computational linguistics /natural language processing (NLP), hypertext/hypermedia/multimedia, etc. Since 2018 he has served as Chief Technology Officer for Mayfair Group LLC, a startup in the legal and insurance fields, and is PI of a related NSF I-Corps study of automated summarization.

Dr. Yinlin Chen holds a Ph.D. in Computer Science and Applications from Virginia Tech, and a M.S. and a B.S. in Computer Science from National Tsing Hua University, Taiwan. He is a Digital Library Architect and Assistant Professor at the Virginia Tech Libraries. His professional interests include Digital Libraries, ML/AI, Cloud Computing, and Serverless Computing. He has participated in and presented his works in conferences and workshops in these areas. He leads a team at University Libraries building the next generation of Digital Library platform and Library services in the cloud environments.

II. Aims, scope, and learning objectives

A. Aims, Scope

This tutorial aims to prepare attendees for the rest of the conference, by giving them an overview of the field, providing pointers so they can share their knowledge with others after the conference (or train or teach others), and connecting them to current technologies and approaches.

It will cover many concepts and use multiple case studies. It will touch upon international activities, including DELOS/DL.org, but drawing in particular on four books about DLs [1]–[4], that build upon the 5S framework, and cover:

- Introduction; Exploration; Evaluation; Integration
- Complex Objects; Annotation/Subdocuments; Ontologies
- Classification; Text Extraction; Security
- Content-based Image Retrieval; Education; Social Networks
- Bioinformatics, eScience, and Simulation DLs
- Geospatial Information

Regarding further DL sharing/training/teaching and learning, coverage includes:

- Overview of the DL curriculum project and methods
- Pedagogical and curricular recommendations
- Discussion of problem-project based learning about DLs

Regarding current technologies and approaches, coverage includes:

- Use of container orchestration with Kubernetes and Elasticsearch
- Use of NLP, deep learning, and ML methods and tools, including for reference analysis, figure extraction, automatic classification, and text summarization
- Use of Jupyter notebooks to conduct ML experiments for training and inference while using the AWS or Google Cloud infrastructure and services.

- Use of managed services and microservices with CI/CD pipeline to construct and deliver DL services
- Use of cloud computing and serverless computing to re-architect traditional monolithic software stacks to serverless and microservices architectures, towards the next generation of DLs.

B. Learning Objectives

Attendees will be able to:

- Explain 5S; compare it with DELOS/DL.org works.
- Describe core DL content/services, informally/formally.
- Describe common DL application areas, from both a user and a system perspective.
- Describe Cloud and Serverless technologies that extend DL capabilities.
- Describe how to handle large tweet, webpage, and book collections, with NLP and ML operations (MLOps).
- Identify modules of interest for teaching or study about DLs, or for use in DL courses, or in courses where DL content can be added.
- Apply problem-project based learning in DL education.
- Explain different serverless architectures and how to apply it in different DL scenarios.
- Apply cloud services and microservices to build DL services.

III. Intellectual / conceptual background

A. Introduction, 5S Framework

The intellectual background for this tutorial is a theoretical framework called '5S': Societies, Scenarios, Spaces, Structures, Streams [4]. Highlights of this tutorial include the applications of digital libraries [1] and the underlying technologies [2], which include: Exploration, Evaluation, Integration, Complex Objects, Annotation/Subdocuments, Ontologies, Classification, Text Extraction, Security, Content-based Image Retrieval, Education, Social Networks, Bioinformatics/ eScience/Simulation, and Geospatial Information.

Supporting all of those are integration methods, along with suitable schemes for evaluation [3].

B. DL Teaching and Learning

Educational resources from an US NSF funded grant to develop DL curriculum (see https://bagua.cct.lsu.edu/dlcurric/) will be presented. Based on the above, discussion of how to learn more about DLs, and how to teach others about DLs, will be tailored to attendee interests.

C. Digital Library on Serverless Computing

Discussion based on the 5S framework will explain how to implement a DL using serverless architectures. Four architectures will be introduced, including microservices, cloud-native, event-driven serverless, and cloud-based architecture. Under the serverless architecture, a DL can continue to evolve to accomplish tasks (Scenarios) under the constraints in local, cloud, or hybrid environments (Spaces). Managed services and microservices are organized as an organic group (Structures) and communicate with each other to perform different tasks (Societies). Tasks flow through different services and each step is triggered by events (Streams). Best practices and demonstrations will be presented.

IV. Tutorial history

Prior related tutorials have been given at: CIKM 95; DL 98-00; ECDL 00, 01, 05-07, 10; ICADL 00-05, 07; JCDL 1-6, 8-11, 13-20; MM 96, 98, 00; SIGIR 96, 01, 05; etc.

V. Target audience, including level of experience

Introductory or intermediate: Those new to the DL field, or coming to it from a different but related discipline, or just new to JCDL, should find this helpful, as they expand their involvement in the DL community.

This tutorial also should be of interest to those already involved in digital libraries, especially if they wish to organize/ solidify their understanding and broaden their perspective, become formally grounded, or to teach a DL course.

VI. Keywords

5S, Societies, Scenarios, Spaces, Structures, Streams, Cloud, Containers, Machine Learning, Natural Language Processing, Serverless

Acknowledgment

Thanks go to NSF, IMLS, National Inst. of Justice, NIH, and QNRF for support through multiple grants. The opinions expressed in this document are solely those of the author.

References

[1] E. Fox and J. Leidig, *Digital Library Applications: CBIR, Education, Social Networks, eScience/ Simulation, and GIS*. San Francisco: Morgan and Claypool Publishers, 2014. [Online]. Available: http://dx.doi.org/10.2200/S00565ED1V01Y201401ICR032

[2] E. Fox and R. Torres, *Digital Library Technologies: Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security*. San Francisco: Morgan and Claypool Publishers, 2014. [Online]. Available: http://dx.doi.org/10.2200/S00566ED1V01Y201401ICR033

[3] R. Shen, M. Goncalves, and E. Fox, *Key Issues Regarding Digital Libraries: Evaluation and Integration*. San Francisco: Morgan and Claypool Publishers, 2013. [Online]. Available: http://dx.doi.org/10.2200/S00474ED1V01Y201301ICR026

[4] E. Fox, M. Goncalves, and R. Shen, *Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach*. San Francisco: Morgan and Claypool Publishers, 2012. [Online]. Available: http://dx.doi.org/10.2200/S00434ED1V01Y201207ICR022