Introducing Computational Thinking into Archival Science Education

William Underwood
College of Information Studies
University of Maryland
College Park, USA
underwod@umd.edu

David Weintrop
College of Education
University of Maryland
College Park, USA
weintrop@umd.edu

Michael Kurtz
DCIC Center
University of Maryland
College Park, USA
mkclandcats@verizon.net

Richard Marciano
College of Information Studies
University of Maryland
College Park, USA
marciano@umd.edu

***Abstract*— The discipline of professional archivists is rapidly changing. Most contemporary records are created, stored, maintained, used and preserved in digital form. Most graduate programs and continuing education programs in Archival Studies address this challenge by introducing students to information technology as it relates to digital records. We propose an approach to addressing this challenge based on introducing computational thinking into the graduate archival studies curriculum.**

Keywords*— *computational thinking, archival science, MLIS curriculum

I. Introduction

The concept of computational thinking is being introduced to STEM education. Wing defines it as a form of problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale [1]. In [2], Weintrop et al propose a definition of computational thinking for mathematics and science in the form of a taxonomy consisting of four main categories: data practices, modeling and simulation practices, computational problem solving practices, and systems thinking practices. In formulating this taxonomy, they draw on the existing computational thinking literature, interviews with mathematicians and scientists, and exemplary computational thinking instructional materials. This work was part of an effort to infuse computational thinking into high school science and mathematics curricular materials. They argue for the approach of embedding computational thinking in mathematics and science contexts, present the taxonomy, and discuss how they envision the taxonomy being used to bring current educational efforts in line with the increasingly computational nature of modern science and mathematics.

The Digital Curation Innovation Center's (DCIC) Computational Archival Science (CAS) collaborative [3] defines computational archival science as [4]:

> *"A transdisciplinary field concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access, with the aim of improving efficiency, productivity, and precision in support of appraisal, arrangement and description, preservation, and access decisions.*
>
> *The intent is to engage and undertake research with archival materials as well as apply the collective knowledge of computer and archival science to understand the ways that new technologies change the generation, use, storage, and preservation of records and the implications of these changes for archival functions and the societal and organizational use and preservation of authentic digital records.*
>
> *This suggests that computational archival science is a blend of computational and archival thinking."*

The argument for integrating computational thinking ideas into archival sciences parallels the case for its inclusion in mathematics and science classrooms. First, archival collections are increasingly comprised of, or at least contain digital materials. This can include native digital entities (e.g. emails, tweets) or non-digital artifacts converted into digital formats for long terms storage (e.g. scans of images). A basic understanding of the characteristics, strengths, and limitations of such computational artifacts is important for future archivists. At the same time that the content is becoming more dependent on technology, so too is the nature of conducting archival work. The tools and practices associated with contemporary archival activities are increasingly dependent on computing. Related to this, the way users interact with archival collections and their expectations of what is possible reflects the increasingly computationally-mediated nature of our world. Collectively, this shifting landscape of archival work means that in order for today's learners to succeed in future archival tasks, it is essential that computational thinking is included as part of their training.

In this paper, we describe our approach to integrating computational thinking into an MLIS program in Archival Studies. It is based on an approach adopted in defining the undergraduate curricula in Computer Science [4]. That approach defined the body of knowledge of computer science in terms of (1) *Knowledge Areas*, (2) *Knowledge Units* within those areas and (3) *Topics* within those knowledge units.

II. Archival Body of Knowledge

The Society of American Archivists states in its Guidelines for a Graduate Program in Archival Studies (GPAS) Curriculum [5] that "Core archival knowledge embraces three separate but interrelated facets of archival studies:

(1) *Knowledge of Archival Material and Archival Functions* (theory and methodology associated with specific areas of archival work);

(2) *Knowledge of the Profession* (history of the profession and evolution of archival practice); and

(3) *Contextual Knowledge* (the contexts within which records are created, managed, and kept"

With regard to *Knowledge of Archival Material and Archival Functions* it is stated:

> *"Archival education should teach the fundamental concepts concerning the nature of archival material in all forms and archival functions (archival theory), the techniques for performing archival functions (archival methodology), and the implementation of theory and method in real situations (archival practice)."*

The components of Archival Material and Archival Functions are:

- The Nature of Records and Archives
- Appraisal and Acquisition
- Arrangement and Description
- Preservation
- Reference and Access
- Outreach and Advocacy
- Management and Administration
- Records and Information Management
- Digital Records and Access Systems

The components of Knowledge of the Profession are:

- History of Archives and the Archival Profession
- Records and Cultural Memory
- Ethics and Values

The components of Contextual Knowledge are:

- Social and Cultural Systems
- Legal and Financial Systems

A description of each knowledge area is included in the GPAS Curriculum. In the terminology of the framework we are developing, the facets will be termed ***Areas of Knowledge***, the components will be termed ***Knowledge Units,*** and we will be seeking to define ***Topics*** within those units.

Our approach to defining ***Topics*** is to:

(1) identify courses in the University of Maryland iSchool MLIS curriculum, and the curricula of other iSchools, corresponding to the SAA core curriculum, along with courses relevant to digital curation;

(2) identify concepts and practices taught in these courses by reference to course descriptions and syllabi;

(3) ask a sample of instructors of those courses to participate in the project confirming topics and providing feedback.

At the same time, the current Library and Archival Science research and practice literature will be reviewed to identify computational concepts, methods and practices that complement the Library and Archival Science concepts, methods and practices already included in the library and Archival Science knowledge base. We call this knowledge base **"The Computational Framework for Library and Archival Education."** We will show how Library and Archival Science curricula can be developed from this knowledge base that include core Library and Archival Science topics as well as complementary computational topics.

Figure 1 shows the correspondences that we have identified between *Knowledge Units* (components) and the University of Maryland MLIS courses [7].

Figure 2 shows an example of what the *topics* and *learning outcomes* might be for the Knowledge Unit *Archival Arrangement and Description*. It was derived the UMD iSchool course syllabus for *INST 782 Arrangement, Description and Access for Archives*.

Knowledge Units	UMD iSchool Graduate School Courses
The Nature of Records and Archives	**INST 646:** Principles of Record & Information Management
Appraisal and Acquisition	**LBSC 785:** Documentation, Collection and Appraisal of Records
Arrangement and Description	**INST 782:** Arrangement, Description and Access for Archives
Preservation	**LBSC 786:** Library and Archives Preservation **INST 784:** Digital Preservation
Reference and Access	**INFM 605:** Users and Use Context **INST 734:** Information Retrieval Systems
Outreach and Advocacy	**LBSC 708W:** Exhibitions, Public Programs, and Outreach in Libraries, Archives and Museums **LBSC 723:** Advocacy and Support for Information Services
Management and Administration	**INFM 612:** Management of Information Programs and Services
Records and Information Management	**LBSC 646:** Principles of Record & Information Management
Digital Records and Access Systems	**INST 647:** Management of Electronic Records & Information **INST 784:** Digital Preservation
History of Archives and the Archival Profession	**LBSC 708P:** Preserving Memory: Archives and Archivists in America
Records and Cultural Memory	**LBSC 708P:** Preserving Memory: Archives and Archivists in America
Ethics and Values	**INFM 612**: Management of Information Programs and Services
Social and Cultural Systems	**INST 643:** Curation in Cultural Institutions
Legal and Financial Systems	**INST 615:** Legal Issues in Managing Information **INFM 722:** Copyright, Privacy, and Security in Digital Information

Figure 1. *UMD iSchool Courses Corresponding to SAA Areas of Archival Knowledge*

Archival Arrangement and Description

Topics:

- Provenance
- Fonds
- Principle of Respect des Fonds
- PrInciple of Respect for Original Order
- Levels of Arrangement
  - Repository
  - Record Group and subgroups
  - Series and subseries
  - File Unit
  - Item
- Finding Aids
- Archival Catalog
  - Series Title
  - Scope and Content Note (range of dates, extent, content)
  - Subjects
  - Record Types (photographs, recordings, Video, digital image)
- Describing Archives: A Content Standard (DACS)
- Encoded Archival Description (EAD)
- Descriptive metadata
- Dublin Core (DC)
- Named Entity Recognition (NER) [9]
- Metadata extraction from text [9]

Learning Outcomes:

1. Explain the principles underlying archival arrangement.
2. Recognize and describe arrangements of records.
3. Describe archival materials according to DACS and EAD.
4. Describe records in terms of Dublin Core metadata elements.
5. Use computational tools for extracting metadata from digital records. [9]
6. Write procedure to identify tokens in digital text documents
7. Write procedures to identify person names, organization names, and location names in text.

Figure 2. *Example of Topics and Learning Outcomes for the Knowledge Unit "Archival Arrangement and Description"*

The Knowledge Units will include indications of core and elective archival science topics and practices. The minimum number of instructional hours for core topics will be suggested.

Named Entity Recognition (NER) and automatic metadata extraction have been introduced as topics and as part of the learning outcomes of the *"Arrangement and Description"* Knowledge Unit. They may not be part of the core, but can at least be elective topics and learning outcomes. A list of references for topics will likely be developed.

Lesson plans will be developed for introducing computational thinking into archival concepts and practices. Exercises and projects that require computational thinking applied to archival tasks will be suggested.

Most MLIS Archival Science curricula are represented as course syllabi. While there are similarities in Archival Science course syllabi across graduate schools, there is no standard set of core topics and learning outcomes. The creation of a Framework with Knowledge Units, Topics and Learning Outcomes that reflect the content of most, if not all, of the major graduate school programs in Archival Science will allow us to be more explicit as to which topics and learning outcomes involve computational thinking. This is one of the primary reasons for developing the Computational Framework in terms of Knowledge Units, Topics and Learning Outcomes.

III. COMPUTATIONAL THINKING

In [2] Weintrop et al identified the following 22 computational thinking practices in science and math education with detailed descriptions of each:

- Collecting Data
- Creating Data
- Manipulating Data
- Analyzing Data
- Visualizing Data

- Using Computational Models to Understand a Concept
- Using Computational Models to Find and Test Solutions
- Assessing Computational Models
- Designing Computational Models
- Constructing Computational Models

- Preparing Problems for Computational Solutions
- Computer Programming
- Choosing Effective Computational Tools
- Assessing Different Approaches/Solutions to a Problem
- Developing Modular Computational Solutions
- Creating Computational Abstractions
- Troubleshooting and Debugging

- Investigating a Complex System as a Whole
- Understanding the Relationships within a System
- Thinking in Levels
- Communicating Information about a System
- Defining Systems and Managing Complexity

Each of these computational thinking practices is described in detail. We show a description of one of them, *Computer Programming*:

> *"The ability to encode instructions in such a way that a computer can execute them is a powerful skill for investigating and solving mathematical and scientific problems. Programs ranging from ten-line Python scripts to multimillion-line C libraries can be valuable for data*

collection and analysis, visualizing information, building and extending computational models, and interfacing with other existing computational tools. This practice consists of understanding and modifying programs written by others, as well as composing new programs or scripts from scratch. This category includes understanding programming concepts such as conditional logic, iterative logic, and recursion vas well as creating abstractions such as subroutines and data structures. While it is not reasonable to expect all students to be programming experts, basic programming proficiency is an important component of twenty-first century scientific programs and use these skills to advance their own scientific and mathematical pursuits." [2]

IV. Computational Thinking in an Archival Science Curriculum

Figure 3 shows a few examples of Archival Science related research and practice literature that involve computational techniques or computational thinking that correspond to Archival Science Knowledge Units. This research and practice literature will be expanded and as topics and learning outcomes are developed for these Knowledge Units, lesson plans and exercises that involve computational thinking will be developed for inclusion in the Framework.

Knowledge Units	References to Computational Thinking and Techniques
The Nature of Records and Archives	Enriched Archival Science concepts and practice through linguistic models and graph theory [10]
Appraisal and Acquisition	Assist appraisal via digital forensics tools, natural language processing, and machine learning [11]
Arrangement and Description	NER & Metadata extraction for Archival Description [12]
Preservation	Digital Preservation [13]
Reference and Access	Faceted Search [14]
Records and Information Management	IDEF0 Model of Records Management [15]
Digital Records and Access Systems	National Archives Online Catalog [16]
Social and Cultural Systems	Data Visualization & Data Analytics [17]
Legal and Financial Systems	Predictive coding in e-Discovery [18]

Figure 3. *Examples on Computational Techniques Related to Topics in Archival Science Knowledge Units*

We envision that computational thinking is integrated into the Masters-level archival studies curriculum by:

(1) creating a course that teaches computational thinking applied to topics drawn from the Knowledge Areas,

(2) introducing computational thinking practices into examples, exercises and projects of graduate courses that cover core archival studies topics.

This doesn't mean that computational thinking practices are integrated into all courses.

V. Computational Thinking in Archival Science in Practice

We will be developing a series of computational thinking lesson plans for graduate school classes in Archival Science. These activities will be informed by the computational thinking taxonomy. The lesson plans can be introduced in professional development workshops to graduate school instructors without formal training in computational thinking or computer science. The lesson plans will be developed so that they can easily be incorporated into existing Archival Science class syllabi. We will be describing them in a more general *Computational Framework for Library and Archival Education*. Here we describe one of the lesson plans that we are developing in the area of Digital Preservation. The contents of the paper in this workshop on "Automating the Detection of Personally Identifiable Information…" [8] also form the basis of a lesson plan incorporating computational thinking.

Preserving Digital Records in a JAR

This activity is designed for a graduate level class in digital preservation and has the students face the task of preserving e-records created using obsolete office automation software. The students are given a Java Archive (JAR) file containing directories of files created on a legacy Windows operating system, using office applications of the 1990s. They learn about file format signatures, the PRONOM file format registry, use DROID to automatically recognize the file formats of the files in the JAR and update the JAR manifest with corresponding file format identifiers. They use a collection of viewers and players to render some of the files. Lacking viewers or players for some of the files, they investigate the use of file format converters to migrate those files to formats for which there is software to render the files viewable or playable. They assess the quality of the converted files, learning about file format preservation standards. They record preservation actions such as the file format conversions in the JAR manifest while learning about preservation metadata standards. They learn to use the features of the Java Archive file to create hash codes for the files to ensure data integrity of the files. The also use hash codes of the information in the manifest to ensure that no e-records are inserted into or deleted from the JAR. Finally, they use a digital signature to sign the JAR, thus learning how these operations can ensure the authenticity of the digital records preserved in the JAR.

VI. Future Work

The Computational Framework for Library and Archival Education will be presented at a two-day Workshop at the U. of Maryland, hosted on Apr. 3&4, 2019 in conjunction with the 2019 iConference. A syllabus for a course on computational thinking for library and archival students will also be

presented. The invited participants will share their insights and experiences to refine the framework.

ACKNOWLEDGMENT

We wish to acknowledge support from the Institute for Museum and Library Science (IMLS) for their support under a Laura Bush 21st Century Librarian (LB21) National Forum Grant [RE-73-18-0105-18] which focuses on planning a workshop of library and archival educators and technologists.

REFERENCES

[1] J. Wing. 2006) Computational Thinking. Communications of the ACM. 49(3), 33-35. Retrieved from https://www.cs.cmu.edu/~15110-s13/Wing06-ct.pdf

[2] David Weintrop, et al (2016) Defining Computational Thinking for Mathematics and Science Classroom. Journal of Science Education and Technology, 25: 127-147. See: https://www.terpconnect.umd.edu/~weintrop/papers/WeintropEtAl_2015_DefiningCT.pdf

[3] DCIC Computational Archival Science (CAS) Portal. See: http://dcicblog.umd.edu/cas/

[4] R. Marciano, V. Lemieux, M. Hedges, M. Esteva, W. Underwood, M. Kurtz, and M. Conrad (2018). Archival Records and Training in the Age of Big Data, in Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education (Advances in Librarianship, Volume 44B, pp.179-199) . Eds: J. Percell, L. C. Sarin, P. T. Jaeger, J. C. Bertot. Emerald Publishing Limited. See: http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2017/06/Marciano-et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final.pdf

[5] Joint Task Force on Computing Curricula, Association for Computing Machinery (ACM) and IEEE Computer Society. Computer Science Curricula 2013: Curriculum Guideline for Undergraduate Degree Programs in Computer Science. ACM New York, NY, USA, 2013, 514 pages. See: www.acm.org/binaries/content/assets/education/cs2013_web_final.pdf

[6] Society of American Archivists. (2016) GPAS Curriculum. See: https://www2.archivists.org/gpas/curriculum

[7] iSchool. Graduate Course Catalog. University of Maryland. See: https://ischool.umd.edu/courses

[8] R. Marciano, W. Underwod, M. Hannaee, C. Mullane, A. Singh and Z. Tethong. (2018) Automating the Detection of Personally Identifiable Information (PII) in Japanese-American WWII Incarceration Camp Records. Proceedings of IEEE Big Data Conference 2018, CAS Workshop, Seattle, Washington. See: http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/10/pii_paper.pdf

[9] W. Underwood, Grammar-Based Recognition of Documentary Forms and Extraction of Metadata, International Journal of Digital Curation, Vol 5, No 1 (2010). www.ijdc.net/article/view/152/215

[10] K. Thibodeau. Breaking Down the Invisible Wall to Enrich Archival Science and Practice. CAS Workshop, Proceedings of IEEE Big Data 2016. See: http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/6.pdf

[11] C. A. Lee. Computer-Assisted Appraisal and Selection of Archival Materials. See: Proceedings this Workshop.

[12] W. Underwood, R. Marciano., S. Laib et al. Computational Curation of a Digitized Record Series of WWII Japanese-American Internment. CAS Workshop, Proceedings of IEEE Big Data 2017. See: http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2017/06/Underwood.pdf

[13] K. Thibodeau. Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, The State of Digital Preservation: An International Perspective. Council on Library and Information Resources Washington, D.C. July 2002, pp. 4-31. See: www.clir.org/wp-content/uploads/sites/6/pub107.pdf#page=10

[14] Daniel Tunkelang.(2009) Faceted Search. See: www.iro.umontreal.ca/~nie/IFT6255/Books/FacetedSearch.pdf

[15] L. Duranti, T. Eastwood & H. MacNeil. Preservation of the Integrity of Electronic Records. University of British Columbia. See: www.interpares.org/UBCProject/index.htm

[16] Research Our Records, National Archives and Records Administration. See: www.archives.gov/research

[17] R. Cox, S. Shah, W. Frederick, T. Nelson, W. Thomas, G. Jansen, N. Dibert, M. Kurtz & R. Marciano. A Case Study in Creating Transparency in Using Cultural Big Data: The Legacy of Slavery Project. See: Proceedings this Workshop.

[18] J. R. Baron, R. C. Losey, and M. D.Berman (eds.), (2016) Perspectives On Predictive Coding and Other Advanced Search Techniques for The Legal Practitioner. American Bar Association.