Heuristics for Assessing Computational Archival Science (CAS) Research: The Case of the Human Face of Big Data Project

Myeong Lee, Yuheng Zhang, Shiyun Chen, Edel Spencer, Jhon Dela Cruz, Hyeonggi Hong, Richard Marciano
College of Information Studies (“Maryland’s iSchool”)
University of Maryland
College Park, United States
{myeong, yzhang63, schen122, espence2, jdelacru, hhong17, marciano}@umd.edu

Abstract—Computational Archival Science (CAS) has been proposed as a trans-disciplinary field that combines computational and archival thinking. To provide grounded evidence, a foundational paper explored eight initial themes that constitute potential building blocks [1]. In order for a CAS community to emerge, further studies are needed to test this framework. While the foundational paper for CAS provides a conceptual and theoretical basis of this new field, there is still a need to articulate useful guidelines and checkpoints that validate a CAS research agenda. In this position paper, we propose heuristics for assessing emerging CAS-related studies that researchers from traditional fields can use in their research design stage. The Human Face of Big Data project, a digital curation and interface design project for urban renewal data, is presented and analyzed to demonstrate the validity of the suggested heuristics. Finally, implications for CAS and future work are discussed.

Keywords—Computational Archival Science; assessment heuristics; urban renewal; data platform; digital curation

I. INTRODUCTION

In April 2016, Computational Archival Science (CAS) was proposed as a new field at the CAS Symposium [2], and subsequently, the first CAS workshop was organized at the IEEE Big Data 2016 Conference [3]. In 2017, this academic movement was articulated as a trans-disciplinary field by combining computational thinking and archival thinking [1]. The working definition of CAS, according to this foundational paper, is as follows:

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 in which 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.

Also, the foundational paper provides eight initial themes that constitute potential building blocks for CAS. These are case-studies at the intersection of traditional computer science and archival science that resonate with the working definition of CAS [1]. Given the approaches and methods that diverse CAS-related studies use, however, it is still unclear what components constitute a CAS study from a researcher perspective. Is the use of computational techniques in an archival science setting sufficient? Do we need some deeper linkages between the two fields? Or is it just a description of an academic trend rather than an active effort to converge ongoing research? These questions can be partially answered by providing heuristics for assessing CAS research.

In order to develop these heuristics, core literature that conceptualizes and defines CAS as a new field are reviewed. Also, other computational fields are discussed to triangulate the position of CAS as an emerging field. Based on reviewing this literature, we propose three protocols to heuristically examine a CAS study: (1) identifying the kinds and depth of computational thinking, (2) blending theories from both fields, and (3) positioning research in and around CAS studies.

It would be challenging to attempt to draw clear boundaries of CAS across many different kinds of computational and archival studies. This position paper is not intended to make distinctions between CAS and non-CAS work; rather, it strives to provide a general guideline or assessment framework that can be used to characterize CAS-related studies so that researchers can get an initial understanding of whether their study is at the core of CAS as a trans-disciplinary field or rather on the periphery.

After proposing our assessment heuristics, the Human Face of Big Data project, a digital curation and interface design study for urban renewal data, is presented and analyzed in order to validate them. This project is concerned with curating close to a thousand legal documents related to
property acquisition in the City of Asheville, North Carolina, during the Urban Renewal period of the 1960’s and 1970’s. The Human Face of Big Data is ideally suited for our validation purpose, since interface design was not one of the foundational CAS paper topics; also, it uses a number of computational techniques to curate archival materials such as geospatial data processing, database design, and visualization. Finally, we discuss the implications of the proposed heuristics on CAS, and the limitations of this approach.

II. RE-THINKING CAS COMPONENTS

A. Computational thinking

The "Computational" term in CAS comes from computational thinking, which is a fundamental skill and a way of thinking that involves multi-layer abstractions, decomposition of a complex task, heuristic reasoning in discovering and solving diverse problems even outside of the field of computer science [4]. This trend has been adopted by many other fields such as biology, physics, chemistry, information science, engineering, and lately journalism, and become an essential part of many disciplines, and it has also claimed that computational thinking is not the identity of computer science, but one of the key qualifications of computer scientists [5]. While some disciplines have successfully established interdisciplinary or trans-disciplinary fields by combining this way of thinking (e.g., computational social science [6]), many fields still remain as traditional fields using extensive methods that are based on computational thinking.

B. What makes an academic field?

It is obvious that computational thinking is a core of CAS as it provides useful skills, tools, and ways of solving archival problems. Is this enough to make CAS a new field? If not, what are the features that lend themselves to a new field? In order to answer this core question, we first need to distinguish between multidisciplinary, interdisciplinary, and trans-disciplinary concepts [7] in order to understand what defines an academic field. According to that paper, multidisciplinarity features the combination of knowledge from different fields but the boundary of each field remains stable. Interdisciplinarity is more about combining form than multidisciplinarity by blurring the boundaries of fields, creating a new way of thinking and its own profession at the intersection of two fields. Trans-disciplinarity is characterized by the recombination of elements from each field creating new knowledge, and by the boundaries of different academic fields being merged.

In the foundational CAS paper [1], CAS is defined as a trans-disciplinary field, and provides justification for going beyond multidisciplinarity and interdisciplinarity. According to the paper, both computer science and archival science use provenance, a predominant means of archival retrieval based on archival administration and reference practices of archivists [8], to capture and assess the trustworthiness of information, which share a common problem [9], [10]. Furthermore, abundance of digital records and the increasing complexity of distributed computing systems motivate the need for an integrated theory by combining knowledge of distributed systems and archival principles [1]. This evidence and socio-technical change justify CAS as an emerging trans-disciplinary field. As a result, computational thinking of archival problems and trans-disciplinarity could be two core qualifications of a CAS study.

C. Positioning CAS Studies

Based on the two qualifications of a CAS study, the capability to map a study onto the landscape of other CAS studies becomes the final assessment stage. When the case studies described in the foundational paper are viewed with respect to the two lenses of computational thinking and trans-disciplinarity [1], they vary from interdisciplinary to trans-disciplinary, while exercising some aspects of computational thinking. It is our intent to contribute to the mapping of the current landscape of CAS-related studies by placing each one on a conceptual CAS spectrum. We wish to help position future studies on a map of current contributions to provide guidance for future research design.

III. THE PROPOSED ASSESSMENT HEURISTICS

An important question for designing heuristics is whether using computational thinking in archival problems is related to integrating or merging two fields, computer science and archival science. As Denning [5] suggested, a big portion of computational thinking has been embedded in and developed with many sciences even before computer science took a major role in advancing this paradigm (just like statistical methods has been dominant tools in many fields). It might be the case that there are advanced sets of techniques, theories, and thinking processes that have been contributed by computer science within general computational thinking practices (e.g., machine learning algorithms vs. task decomposition in management [11]).

This suggests that using computational thinking to solve archival problems does not necessarily indicate that computer science and archival science are merging. Rather, it is possible to describe the relationship between computational thinking and the degree of integration between the two fields as being orthogonal where each component creates a spectrum. In other words, computational thinking can create a spectrum where one end shows a high degree of using advanced skills and thinking processes, while the other end indicates general, non-computer-science-specific practices. In a similar way, the extent to which two fields are merged creates a spectrum where one end is multidisciplinarity and the other end is trans-disciplinarity (interdisciplinarity is in
between). Figure 1 depicts this model and relationship as a perceptual map.

Many studies in traditional archival science have already been using parts of computational thinking as general practices, so it is possible to view this discipline as being already at the stage of interdisciplinarity while varying to some degree. There are increasing number of studies [1] that use advanced-level computational thinking processes and methods, which can place each study on the spectrum of collaboration and core CAS. If a study uses advanced-level computational methods such as Natural Language Processing (NLP) and machine learning, but does not create shared goals and its own profession in the problem space, it can be characterized as a multidisciplinary collaboration. If an archival and computational problem starts creating shared goals, its own profession, and new theories, it could become a trans-disciplinary study, i.e., a core CAS study.

Of course, there are no clear boundaries between these characterizations, and this map provides an initial and abstract understanding of where each study could be placed. Some case studies introduced in the foundational paper can be also presented on the map. For example, the Open Archival Information System (OAIS) reference model was developed at the intersection between computer science and archival science to increase the value of preservation [12]. This model consists of three stages where documents are prepared to describe the state of the data: Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP). Each package contains particular digital artifacts such as technical and administrative or description metadata. This modeling follows a method that is used for packaging software modules, while informs computer science with new types of data artifacts. This characteristic allows us to place this study somewhere between core CAS and collaboration between the two fields on the perceptual map, since it goes beyond a simple collaboration while does not quite reach the stage of creating a new theory or shared goal. Overall, our assessment heuristics appear to be three-fold, and are presented in Table I.

### IV. A CASE STUDY: THE HUMAN FACE OF BIG DATA

In order to test the validity of the proposed heuristics, we assess the Human Face of Big Data project, because this project has unique components such as interface design methods and value-sensitive design, which are not introduced as case studies in the foundational CAS paper.

#### A. Background

From the early 1960s until the mid 1970s, urban renewal projects were implemented in the city of Asheville, North Carolina. Urban Renewal was a nationwide program with the aim of transforming “blighted” neighborhoods into modern housing areas with up-to-date amenities. Communities, in mostly African-American neighborhoods, were disrupted as thriving families, businesses and organizations were displaced in the name of the economic development and infrastructure improvement. Fulfilove describes this as the “traumatic stress reaction to the loss of some or all of ones emotional ecosystem” and she conceptualized this kind of displacement as root shock [13].

The Human Face of Big Data project focuses on records from the East Riverside neighborhood (now referred to as Southside) which was the largest area in the southeastern United States affected by urban renewal. In 1966, there were about 4,000 people and nearly 1,300 households living in this neighborhood. These records are currently stored in the D.H. Ramsey Library Special Collections and University Archives at the University of North Carolina, Asheville. The collection contains the records of the Housing Authority of the City of Asheville (HACA) and is comprised of nearly 130 linear feet and 129 cartons of documents. It includes

<table>
<thead>
<tr>
<th>Table I: Heuristics for accessing CAS research</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heuristic</strong></td>
</tr>
<tr>
<td>Identifying the kinds and depth of computational thinking</td>
</tr>
<tr>
<td>Blending theories from both fields</td>
</tr>
<tr>
<td>Positioning research in and around CAS studies</td>
</tr>
</tbody>
</table>
Figure 2: A multi-layer representation of georeferenced maps on top of the online map (from the top, the 1973 map, the inverted 1966 map, and the modern parcel map).

the legal documents of the acquisition process for nearly 1,000 properties in the Southside neighborhood that were part of the urban renewal experience. The page count of each case file ranges from 20 to 200 and covers the entire sales history of each property. It begins with the appraisals and moves through the offers made, cases of rejection by former residents and the ensuing court cases, and ends with the final deed transfer to HACA.

B. Digital Curation and System Design Processes

A property acquisition document pertains to a property parcel; this led us to begin with the digitization of a parcel map from 1973. After the locations of parcels during the urban renewal period were correctly identified, geo-polygons for these parcels were generated to provide access points to corresponding properties acquisition data. Based on this digitized map, the scenario-based method was used to design user interfaces and databases [14].

1) Digitizing a Map: Geo-referencing and Geo-tracing Properties: Since the original urban renewal map created for the planning purpose in 1973 was very large (over 4 sq. meters), it was scanned in pieces using an A3-size scanner. This resulted in 36 distorted raster images of the map and made it difficult to integrate them back to a whole neighborhood map due to mismatched roads and parcels between each tile and varying levels of shading. To resolve this, Photoshop, particularly the smart objects function, was used to adjust each image, to make streets and parcels between partial images continuous between each other, and to stitch them together. Then, ArcMap, a desktop version of ArcGIS1, was used to geo-reference the merged map. In other words, the integrated raster map was adjusted and rubber-sheeted so it could be correctly pinned onto the modern online map. Since the road structure of Asheville has changed significantly, it was challenging to identify control points that could be used to align the two maps during the geo-referencing process. This issue was resolved by using another layer of a historical map, a 1966 aerial map, as a means to bridge points between the modern parcel map and the 1973 map. In other word, we used the modern parcel layer to geo-reference the 1966 map, and this 1966 map was used subsequently to geo-reference the 1973 map. Figure 2 depicts the multiple layers of maps used.

After the raster map from 1973 was correctly geo-referenced, old parcels were geo-traced using QGIS2. Since there were no digitized polygons that represented each parcel before renewal, it was the only way to manually draw the polygons following the parcel boundaries of the 1973 map. After geo-tracing all the parcels, we exported the geospatial layer as a GeoJSON file. This digitized polygon data was used to provide entry points to users on the online map.

2) User Persona and User Scenarios: The user interface that is based on the urban renewal map of 1973 was designed via the scenario-based approach that consists of two steps. The first step is to create user personas by making an educated guess at which stakeholders would be potential key users of the system. For example, the previous property owners and HACA might be potential users since they are the witnesses of history and participated in historical events. We also tried to figure out other types of potential users who could be motivated to use the data based on the contents of the data and digital map; these potential users included researchers, system designers, and archivists. In total, we generated about 15 personas through multiple brainstorming sessions. Table II shows examples of user personas.

After brainstorming and identifying key stakeholders, we created user scenarios that describe the context and reasons why and how those stakeholders would want to use the system in the future. Three to five user scenarios for each user persona were generated. Table III is an example of user scenario. In this example, some of previous owners might want to use this system to know the market value of their former property in Asheville. Historians might want to use the system for researching the social change that occurred as

1http://www.arcgis.com/
2http://www.qgis.org/
Table II: Examples of user personas

<table>
<thead>
<tr>
<th>Role</th>
<th>Story</th>
</tr>
</thead>
<tbody>
<tr>
<td>[FR1] A former resident who lost his/her house and has not kept in touch</td>
<td>He is in his late 60s and works for a small company with a spouse and two children who go to school. He lived in Asheville for 10 years before they moved out. His family was required to leave their house and move to another city that was miles away, and had no detailed knowledge of what happened to their former house. He had to leave his job in Asheville to move to a new place.</td>
</tr>
<tr>
<td>[AR] An archival researcher</td>
<td>She is an assistant professor from a public university in American Studies. She is a native English speaker in her late 30s. She has experience working at the National Museum of American History (Smithsonian) as an archivist, and has a specialization in integrating historical data from different institutions to provide easy-to-access platforms. She is very knowledgeable in archival science terms, and interested in developing research archival science questions.</td>
</tr>
</tbody>
</table>

Table III: An example of a user scenario

<table>
<thead>
<tr>
<th>Role</th>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>[FR1] A former resident (Male, 68 yo.)</td>
<td>Scenario 1: FR1 search for an address/street name to zoom into his/her former house and see how much it costs. Scenario 2: FR1 wants to know if their property was bulldozed or fixed (partially identifiable). Scenario 3: FR1 wants to know how many days it took to receive compensation and finally give up their house / apartment.</td>
</tr>
</tbody>
</table>

Each user scenario was converted to user requirements that were to be used in designing user interfaces and database structure. For instance, if a previous owner wanted to know the property acquisition process for his or her house, the user can explore this by accessing the system through a web browser, searching for their former address in Asheville, and reading the legal events chronically. These user requirements are connected to actual system design (e.g., the searching functionality is implemented with search boxes, search buttons, a result view, and database queries).

3) Data Modeling: Even though this scenario-based, top-down approach to collect user requirements is a useful way to design user interfaces, the context of historical data platforms informs that the design can be heavily dependent on the contents of archival documents themselves. Furthermore, human stories extracted from the legal documents could affect the system design significantly as they can highlight important values that were overlooked during the urban renewal process. This motivated us to qualitatively examine the original documents and to identify data patterns that could potentially indicate meaningful but previously unknown details. For example, there were instances that when the property owner refused to have their home evaluated, appraisers examined it only by its appearance and valued improvements as zero. A list of these qualitative patterns was created to guide the collection of data from urban renewal documents.

This list of data patterns, particularly that focus on human stories, became another source of system requirements. For example, tenant names and their moving-out dates showed up in the legal documents. Since former tenants can provide human stories behind the community displacement beyond the legal considerations, tenant information such as moving-out dates and rental prices became important part of stories that could benefit potential users.

4) Interfaces and Database Design: Based on our user scenarios and data patterns, wireframes were generated for communicating the user interface design. During the wireframing stage, we tried to assess the types of data that should be displayed on the interface. The final version of hand-drawings were re-generated as a raster image using hand-drawings for interface design.
Photoshop (Figure 3). This process helped us understand the future layout of the interface and the way the system would work.

In addition, a relational database was designed and implemented iteratively by extracting variables from each use case. At first, we tried to store all kinds of entities that showed up in the documents. Due to the lengthy and unstructured nature of the legal documents, however, we found it challenging to make a generalizable database. This led us to re-model the relational database using an event-based, spatio-temporal data model [15]. The database schema was revised several times, and finally implemented as an actual database (Figure 5), using PostgreSQL, an open source relational database system. Due to the inconsistent and unstructured nature of the documents, data was crowdsourced iteratively by a team of researchers.

Finally, the prototype system with a part of the dataset stored in the database was implemented by using PHP, JavaScript, Leaflet[^3], PostgreSQL, and D3[^4]. The interface consists of an online map and a sidebar as shown in Figure 4. The map of Asheville shows the lifecycle of parcels going through acquisition processes, and at the same time, it visualizes the property acquisition process by color-coding parcels while the "year bar" on the left hand side changes. The sidebar on the right provides search functionality based on street name or person name, and a pie chart shows the city-wide status of property acquisition for the selected year.

Users are able to search for a parcel based on part of an owner’s name, tenant’s name, or a street address. A list of parcels that match the input query is displayed at the bottom of the sidebar. When a user clicks a parcel from the list, the corresponding parcel is zoomed in and highlighted on the map. Moreover, each parcel on the interactive map is a polygon, and all polygons are clickable. When a user clicks on a polygon, a pop-up window provides a synopsis of the life of that parcel during the urban renewal period, displaying photos, a legal event timeline, and the individuals associated with that parcel (appraisers, owners, tenants, and purchasees).

[^3]: http://leafletjs.com/
[^4]: https://d3js.org/
5) **Design Process**: The next step is to receive feedback from potential users, so that we can modify user requirements and iteratively refine the system design. This is known as the *iterative design process* [16], and our version of this process is shown in Figure 6. Also, unlike other data platforms, the urban renewal data system is characterized as a historically-sensitive data platform since there are tensions and diverse perspectives between potential stakeholders. This motivated us to use the value-sensitive design (VSD) method in the next iteration of the design process through a set of interviews that will elicit diverse uses [17]. We expect that several iterations will be required and that this will help assess additional dimensions of values and user opinions.

### C. Heuristic Analysis

Since the Human Face of Big Data project makes use of methodologies and thinking processes from other fields including computer science, it is not trivial to anticipate where this project fits regarding CAS. This is the motivation for us to propose a set of heuristics to assess its potential contribution and standing from a CAS perspective. These form an assessment prism and include: computational thinking, theoretical blending, and positioning within the realm of CAS themes.

1) **Identifying the kinds and depth of computational thinking**: The Human Face of Big Data project uses computational thinking processes such as iterative design and agile development, which partially stem from software development practices [18]. Also, it uses data visualization, data management through relational databases, and event-based modeling by computationally reorganizing unstructured archival data [15]. Value-sensitive design (VSD) and scenario-based design are used as main design methodologies. Much of this work has been developed in the human-computer interaction (HCI) field [14], [17]. While HCI is itself another trans-disciplinary field combining many fields, part of the thinking processes and modeling methods still rely on advanced computational thinking. In this sense, the Human Face of Big Data project can be placed on the upper part of the perceptual map, i.e., using more than general-level computational thinking processes.

2) **Blending theories from both fields**: The approaches and skills used in this project contribute to some domains of HCI and agile practices in the context of the digitization of geo-tagged, archival data. Furthermore, design processes developed in the context of historically sensitive data platforms are not only for contextualizing existing VSD theories, but also potentially to contribute to the VSD theories by complementing the less studied body of target systems. This implies that the Human Face of Big Data project makes use of computational thinking, as well as creates a new body of knowledge by combining archival concepts such as preservation, accessibility, models of VSD, iterative design processes, and agile processes. Thus, this project can be placed closer to trans-disciplinarity than multidisciplinarity.

3) **Positioning the Human Face of Big Data Project within the realm of CAS themes**: While the case studies introduced in the foundational CAS paper provide initial CAS reference themes, the case of the Human Face of Big Data suggests that design process, VSD, and event-based data modeling can be potentially added as new CAS themes. In particular, interface design processes and VSD have been less highlighted in archival science maybe due to the importance of dominant archival values such as preservation, authenticity, and accessibility.

Related themes to interface design could be *evolutionary prototyping* and *digital curation*. Several studies mentioned how user feedback was handled in designing and refining prototype systems for archives. For example, The University Libraries at UT Austin created curation tasks which were based on user needs such as expanding the dataset content scope by increasing its information elements, deleting personal data, and repackaging data to make it more accessible and easy to analyze [19]. While prototyping the system, user feedback played an important role in increasing system reusability for diverse datasets.

Another prototype named "The Presidential Electronic Records PiOt System (PERPOS)" was developed for archival processing of digital email records [20]. The main goal of the project was to identify and accurately classify millions of historical email records. In this project, a new method was developed based on an existing algorithm for automatically annotating semantic categories and implemented as a prototype. This gave designers an opportunity to evaluate the system for both functional and non-functional parts. It also guided designers to better understand the system requirements and helped them figure out the approaches to achieve the goal. In addition, the software was able to prove that it could help discover new functional requirements through user feedback. These prototyping processes and user feedback for understanding user requirements and intermediate feedback are at the core of software engineering and computational thinking, and philosophically part of the agile development process [18].

The digital curation theme emphasizes aspects of how to
add value to scientific and humanities data by exercising diverse archival practices such as adding metadata, indexing, and visualizing data. This can be a higher level theme that partially covers prototyping and design processes. However, iterative design and VSD for interfaces have a unique place in CAS. They emphasize more of *what and how* diverse user values could be considered and integrated into the interface design (e.g., user boundaries, ethics, and transparency), rather than focusing on archival values themselves. It is more about humanizing and specializing archival systems rather than generalizing them. We argue that these characteristics make iterative design and VSD unique CAS themes.

Finally, the *spatial and temporal analytics* theme is also relevant to the Human Face of Big Data project as this project leverages spatial and temporal data to store and visualize archival information. Also, geo-referencing and geo-tracing techniques were used to accurately curate a historical map on top of the online map. From the spatio-temporal perspective, this project does not yet reach transdisciplinarity. While it makes use of computational skills in modeling and curating historical data, these practices do not necessarily create a new problem space from a computer science perspective.

V. DISCUSSION

Using the proposed heuristics, it was possible to assess where a digital curation study could be situated on a CAS spectrum. While our perceptual map (Figure 1) provides a useful lens to look at a CAS-related study, parts of the heuristics still need additional articulation and refinement.

One of the complexities encountered in our case study is that computational thinking cannot be easily modeled as a one-dimensional spectrum due to the multiplicity of computational skills and processes. While the design process and VSD portions of the Human Face of Big Data project can be viewed as advanced-level computational thinking that could create a new theoretical basis, other parts of the study such as spatio-temporal analytics and visualization still remain in the multidisciplinary spectrum. One approach would be to assess a study by primarily looking at the most advanced parts. However, our observations suggest that CAS-related studies need to be evaluated from diverse perspectives of computational thinking. In other words, the degree of computational thinking in a study cannot be judged by one or two methods or approaches, but instead involves a multi-dimensional examination to understand it properly. We suggest that a possible approach to map a study is to represent it in multiple perceptual maps where each map expresses a particular aspect of computational thinking and skills. Potential representations are shown in Figure 7.

Another limitation of this approach is that it is not trivial to distinguish between general-level and advanced-level computational thinking processes and skills. To illustrate this point, a few questions can be asked: is the event-based spatio-temporal model advanced or not as a computational technique? Is it less advanced because it was studied mainly in the Geographic Information Systems (GIS) field (which is already computational) rather than computer science? It is generally very challenging to assess the relative level of computational thinking and techniques even among similar qualifications. This might be due to the close but complex relationships between computer science and other fields. This complexity implies that the spectrum of computational thinking needs to be modeled more precisely and further conceptualized.

Despite these imprecisions, it was possible to map a CAS-related study on the CAS spectrum, and to provide a reference point of how it is situated in this emerging field. This heuristic is expected to provide researchers in and around CAS with a useful tool and framework for assessing their studies, especially in their study design stage.
VI. Conclusion

This position paper strived to articulate the meaning and implications of the emerging field of Computational Archival Science (CAS), by integrating computational and archival thinking. As a means to articulate CAS, we proposed heuristics for assessing future CAS research. By using a top-down approach, it was possible to develop three protocols that heuristically examine emerging CAS research. While still a formulation in progress due to limitations in the spectrum model requiring additional validation, the case study of the Human Face of Big Data project provides initial validity of the heuristics. It is our expectation that CAS can be more precisely articulated in the future through the refining of these assessment heuristics by taking into account more diverse case studies and more precise models. As such, we hope this position paper is a contribution towards the formalization of this field.

ACKNOWLEDGMENT

We wish to acknowledge ongoing support from Maryland’s iSchool to the Digital Curation Innovation Center (DCIC) and also for major funding from the National Science Foundation’s “Brown Dog” project (NSF Cooperative Agreement ACI-1261582), a $10.5M NSF/DIBBs-funded collaboration with the University of Illinois at Urbana-Champaign NCSA Supercomputing Center (PI, Kenton McHenry). Brown Dog aims to help accelerate the development of digital curation processes and services and create a data observatory to provide access to Big Records training sets and teach students practical digital curation skills. The Human Face of Big Data has leveraged the iSchool’s cloudLab, and serverFarm as well as the Brown Dog funded dataCave and DRAS-TIC open-source platform (Digital Repository At Scale - That Invites Computation). See: http://dcic.umd.edu/about-us/infrastructure/.

REFERENCES


