Society of American Archivists
COMMITTEE ON RESEARCH, DATA AND ASSESSMENT

Date: 20 April 2020

To: SAA Council

From: Paul Conway & Jennifer Gunter King (CORDA co-chairs)

Subject: CORDA status report

This status report covers the most important activities of the three subcommittees working in tandem under our leadership of CORDA. The report's particular emphasis is on CORDA's plan to create a data repository for the Society of American Archivists and the archival profession at large utilizing the open source Dataverse software platform developed at Harvard University.

Repository subcommittee

CORDA recommends that SAA create a data repository service utilizing the Dataverse open source software and that SAA establish a working relationship with the Odum Institute at the University of North Carolina via a formal Memo of Understanding. CORDA requests that SAA Council make the following three decision regarding the repository project:

Action items for SAA Council, May 2020.

1. Approve collaborating with the Odum Institute at the University of North Carolina to establish and support a pilot instance of Dataverse. Our expectation is for a project of five or six years in duration followed by an explicit decision by Council to extend the service.
2. Authorize the Executive Director to negotiate a Memo of Understanding regarding the rights and responsibilities of SAA and the Odum Institute for deposited data, liabilities for data reuse, and technical support.
3. Reaffirm the responsibilities of CORDA to populate the SAA Dataverse with datasets and establish policies governing contributions and data use.

What follows is background information to support these recommendations.

Why should SAA build a data repository?

In establishing CORDA in late 2018, the SAA Council explicitly charged the new committee with "providing a repository or portal for data and other research outputs." Further, the CORDA charge expects that the Committee will provide SAA members with tools for gathering and analyzing data, training on data analysis, and summaries and synthesis of "facts and figures" about archives and archivists.

> *In short, CORDA is proposing a data repository service because SAA Council charged us to do so and because it is most assuredly the right project for SAA to embrace.*

A data repository would provide a new service for members of the Society of American Archivists and, potentially, for the archival profession at large. As fully envisioned, a repository for SAA would assemble, document, and make available via search tools and download capabilities datasets that have been created by SAA and its organizational units, by regional archival and allied professional associations, by individual archival programs, and by archivists. Data sets could take the form of structured social-science-like data (e.g., surveys) or well organized qualitative data gathered through systematic research or assessment (e.g., interview transcripts, open-ended responses, etc.).

The three truly distinguishing features of a robust data repository are 1) full and complete documentation on the datasets [i.e., codebooks and reports]; 2) cross-searching across the metadata for multiple datasets [i.e., aggregation]; and 3) the possibility of joining data from the repository with easy to use analytical tools and the guidance needed to analyze the data.

A data repository is fundamentally about fostering data reuse for purposes of deepening our understanding of SAA's activities and contributions over time, the status and value of archival organizations, and the evolving professional status of archivists. The types of datasets envisioned for an SAA-managed repository are national level surveys of archivists and archival organizations (e.g., A*Census I, A*Census II, past surveys of archivists or archival organizations), data gathered by SAA sections and roundtables, data initially gathered by large archival organizations that likely has broader uses, and research data gathered by archivists as part of their work or studies.

SAA's data repository is not necessarily about long-term preservation, although the careful choice of a trustworthy repository partner may further assure SAA's digital curation and preservation goals. Instead, an SAA repository is about making an impact through data, information, and knowledge. A data repository will further advance SAA's third strategic goal of "Advancing the Field," particularly 3.2.: "Foster and disseminate research in and about the field." The single biggest measure of success for an SAA data repository is that it is used by SAA, archival organizations, and archivists to further their knowledge and increase their professional and societal impact.

In proposing a pilot repository, likely followed by a limited-term repository project of up to five additional years, CORDA is making a commitment to the success of the repository and to the assessment and evaluation of its value, such that SAA can decide whether to continue, modify, or suspend the service.

Why Dataverse?

From the Dataverse Project website: See: https://dataverse.org/

> "Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data. It facilitates making data available to others, and allows you to replicate others' work more easily. Researchers, journals, data authors, publishers, data distributors, and affiliated institutions all receive academic credit and web visibility. … The central insight behind Dataverse is to automate much of the job of the professional archivist, and to provide services for and to distribute credit to the data creator."

The Dataverse Project is a worldwide consortium of committed repository software developers and open source advocates based administratively at Harvard University. A Dataverse repository is the software installation, which then hosts multiple virtual archives called Dataverses. Each dataverse contains datasets, and each dataset contains descriptive metadata and data files (including documentation and code that accompany the data). The software itself is maintained and upgraded by the open source software programmers in the Dataverse community.

Before Dataverse, researchers were forced to choose between 1) controlling distribution of their data individually without long term preservation guarantees but receiving credit for their work, or 2) sending their data to a professional archive that provided long term preservation but without receiving much credit. Dataverse provides the best of both worlds by embedding a virtual archive on the SAA website that generates the look, feel, branding, and URL of SAA, along with an academic citation for the data that gives the creators of the dataset full credit and web visibility.

The Dataverse platform at the Odum Institute at University of North Carolina (UNC) is one of several database management tools that the Institute uses to acquire and provide access to social science research data. The Odum Institute itself is world-class social science data repository that functions as a semi-independent unit of the university. The Odum Institute was founded in 1924 and has as its mission "to foster groundbreaking social research that improves the lives of people … around the world." UNC considers Odum to be mission critical because of the data repository services that it provides to UNC faculty, staff, and students. CORDA is recommending that SAA partner with the Odum Institute at UNC because their mission and our needs are in sync. See: https://odum.unc.edu/

CORDA is proposing that SAA establish its own Dataverse and simply name it: SAA Dataverse.

What alternatives did CORDA consider?

As one of its first tasks in 2019, CORDA reviewed the capabilities and pros/cons of four robust repository platforms that could possibly host a SAA repository service. The four platforms considered were the Humanities Commons of the Modern Language Association, ALAIR from the American Library Association, LIS Scholarship Archive (LISSA) from the Center for Open Science, and, finally, the Dataverse Project developed by Harvard University and the Odum Institute at UNC, one of 47 instances in a global consortium. Our comprehensive review included assessing technical capabilities, organizational support, file formats and metadata accepted, and collection/access policies.

CORDA also worked through the feasibility of asking SAA to develop and support its own dedicated data repository, somehow associated with SAA's website or publishing platforms. Both of these options proved untenable due to the lack of financial and staff resources, long startup times, and the absence of adequate technical expertise. By building its own data repository, SAA would also miss out on the opportunity to have data associated with other social science research.

CORDA considered but dismissed the option to partner with the Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan. ICPSR is the premier social science data repository in the country and curates datasets from a wide variety of organizational and personal sources for its 750 member organizations. ICPSR assesses a download fee of $550.00 per dataset to non-members and individuals. CORDA member and SAA Fellow Nance McGovern arranged a **one-off** deposit agreement with ICPSR for the A*Census I data in 2006, when she served as preservation archivist for ICPSR. It is CORDA's position that certain vital datasets, such as A*Census II and other national surveys of archival organizations, ought to be deposited with ICPSR with a non-exclusive agreement because of its exquisite data curation processes. SAA should then rely on the Odum Institute and Dataverse to hold other datasets for as long as they are useful.

Costs and Ongoing Support: The Odum Institute assesses no fees to SAA for establishing and supporting the SAA Dataverse. The Odum Institute has conveyed to CORDA that supporting the types of data on archival programs and services is fully in keeping with the core mission of the Institute.

After meeting with the Odum Institute Assistant Director for the Archives, CORDA determined that SAA's Dataverse could be created and sustained using their self-service model, which has a cost of $0 to SAA. This no-cost option is based on the fact that CORDA will not need their assistance with creating metadata templates, organizing and uploading datasets, and other activities for which Odum assesses fees. Odum Institute staff have assured CORDA that is it highly unlikely that the size and scale of the datasets deposited by SAA during the pilot period will stress the server or storage capacities of the University of North Carolina. See the chart of Odum Institute Service Fees at the end of this report.

CORDA expects that launching, branding/customizing the interface, and integrating the SAA Dataverse into SAA's web presence will require approximately 60 hours of effort over one year by SAA staff responsible for SAA's website and marketing program. We do not expect a greater level of staff commitment per year over the life of the project. CORDA can create and maintain the policy infrastructure for the repository and to provide training and guidance on contributing datasets to the SAA Dataverse.

During the pilot phase of the SAA Dataverse and into a sustained period of growth, CORDA will pay very close attention to the number of data deposits and provide ample early warning about storage constraints at Odum. In a way SAA can only wish that the SAA Dataverse is so successful in accumulating relevant and useful datasets that SAA may have to consider supporting Odum's storage needs. This storage condition is highly unlikely to develop for three to five years.

Memo of Understanding: CORDA wishes to work with the SAA Executive Director to negotiate a Memo of Understanding (MoU) between SAA and UNC (Odum Institute) to cover the activities of the pilot project and provide for a smooth transition to an ongoing repository capability if SAA decides to do so.

The purpose of a MoU is to make explicit the responsibilities of UNC for maintaining the Dataverse software platform and access to any data deposited in this repository. The MoU also specifies the responsibilities of SAA to obtain permission for data deposits, waives fees for data deposits and maintenance, releases UNC from legal liabilities relating to deposited data, and provides for termination of the agreement.

CORDA proposes that SAA adapt a similar MoU between UNC and Emory University as a template for an agreement between SAA and UNC. The MoU should be finalized as soon as possible but no later than September 2020.

Timeline for a one-year pilot project

Spring 2020. Complete memo of understanding; brand SAA Dataverse; populate the pilot repository with data from A*Census I and up to five other datasets as appropriate and available.

August 2020. Launch pilot SAA Dataverse as part of annual meeting pre-conference workshops and other scheduled meetings.

Fall 2020. Establish collection and use policies through transparent process with SAA.

Winter 2021.	Prepare and vet technical support plan for SAA Dataverse. Prepare 3-5 year budget proposal for transforming a pilot project into a sustainable initiative.
May 2021.	SAA Council reviews recommendations for ongoing support for SAA Dataverse.
Summer 2021.	Evaluate and assess the viability of the SAA Dataverse and recommend next phase.

Other CORDA Activities

Education and Training subcommittee

The subcommittee is planning two workshops, an open forum, and a special session at the Research Forum for the annual meeting. A half-day limited enrollment workshop teaches archivists how to analyze data from surveys and other quantitative sources and produce appropriate statistics, charts, and graphs. An "unworkshop," open to all meeting participants, will solicit ideas on the types of datasets that archivists have or will find useful in understanding the status or archivists, the functions of archival organizations, and the impact of archivists, archival work, and archival collections in the wider society. A CORDA open forum is designed to update SAA members on CORDA activities and solicit ideas for the Committee.

Beyond these three activities, CORDA is also partnering with the SAA Research Forum organizers to hold a special session at the Forum. The focus of the 60 minute program is a moderated discussion of what professional research and evaluation is, how to define important questions for investigation, and what qualitative and quantitative data CORDA/SAA needs to collect for a repository that would support the desired research.

In the summer of 2020, prior to the SAA annual meeting, the CORDA education and training subcommittee plans to develop and offer one to two hour webinar for SAA leaders on the benefits of a data repository for SAA sections and roundtables.

Assessment and Evaluation subcommittee

This subcommittee's work is focused on identifying existing datasets that could be deposited into the SAA Dataverse, as well as needs for data that could be gathered by SAA groups, by regional archival organizations, or by individual archivists and archival institutions. The first of a number of "shaking the tree" activities were begun through a survey, addressing the CORDA mandate to facilitate research and evaluation activities.

In January of 2020 the subcommittee developed a survey for SAA Section Leadership, putting forward the vision of that CORDA was established to fulfill, and asking for SAA leadership assistance with identifying data and research. The survey was sent to the more than 100 chairs and cochairs of SAA committees and subcommittees.

CORDA Email to SAA Committee Leadership:

> "SAA's Committee on Research, Data and Assessment (CORDA) needs your help. Established last year, CORDA is responsible for providing access to significant and useful data and research about SAA, American archives, archivists, and their users. In addition to developing a research agenda for the profession as a whole, over the course of the coming year CORDA will launch a

> data repository for institutions, groups, and individuals to upload and access relevant research, reports, and raw data. The repository will include a clear collection development policy identifying what we have deemed appropriate and relevant for collection. We want to create a policy and repository that best suits our professional needs, and that's where you come in.
>
> Today we are reaching out to you with an invitation submit, in your capacity as section leaders, information and materials that might be suitable for review for submission to the SAA Research Data Repository. We are looking for any known evaluation and assessment work or research data that you are aware of and including data that your committees and sections might have access to or may have commissioned, that would have potential value to the profession."

CORDA Section Leadership Survey

The survey response is promising and inspiring. We have several responses that include great detail about research initiatives in several areas of the field, including records management, web archiving, and public libraries. The responses include both data and data collection tools. These responses will begin to flesh out the scope of available data as well as the scope of needed research and data. But to do that, we also need a place to put the data.

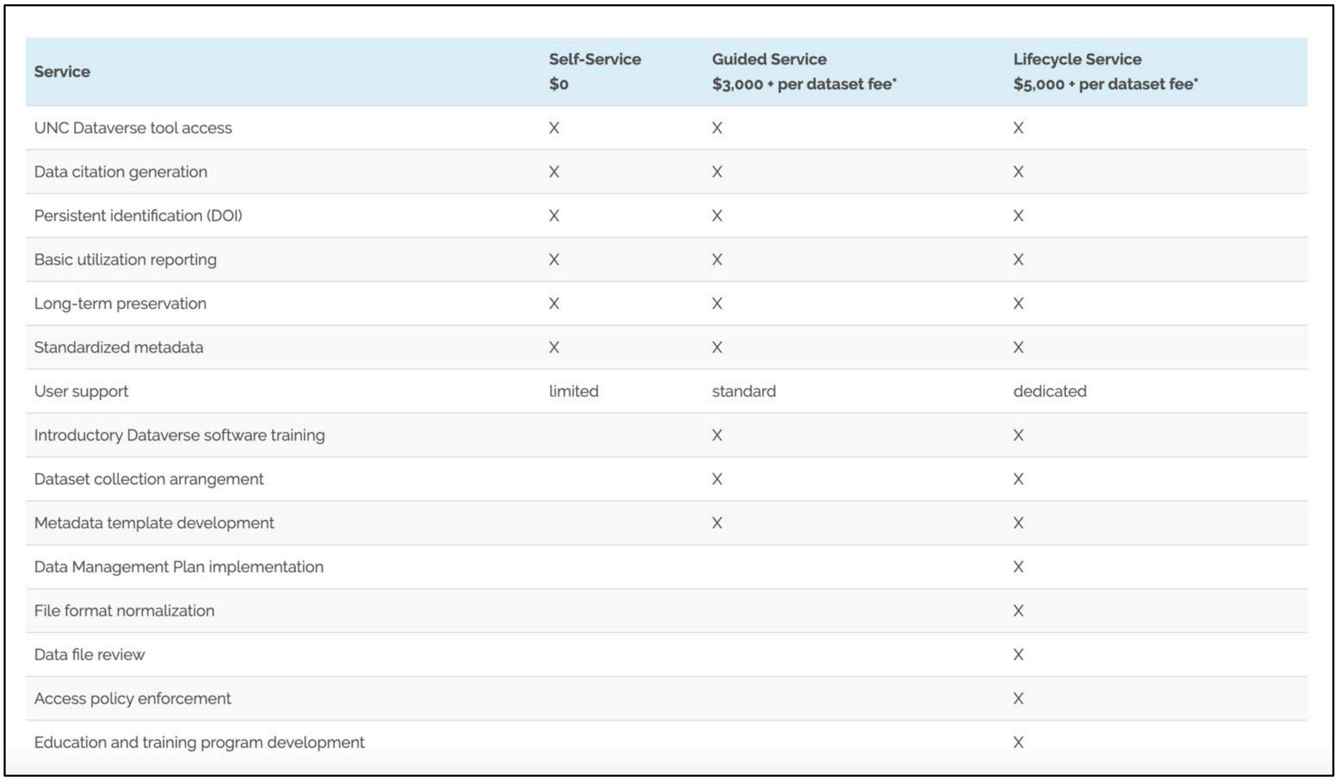
The invitation to SAA leadership to contribute to a broader research environment in the profession is designed to both illuminate research and data that already exists, and also to identify research that is needed. To fulfill this invitation, it is most important that CORDA be able to take the work that Leadership has done through the survey and use the data to start to create a framework as well as present the data in a repository. The repository is a critical step in making visible the importance of research and data for the profession, and without a repository, the work of CORDA and the impact CORDA can have on behalf of SAA will begin to sound like conversation rather than action.

The repository is the cornerstone to the contextual information that CORDA will develop, including the "Facts & Figures" page of the SAA website. Our goal is to assemble information that advocates for the value and impact of the archival profession as well as promotes tools for gathering and analyzing data about archivists and archival organizations.

Next Steps: With the repository established, the survey concluded, CORDA will outline steps for data and collection development policies, and will begin building a robust web page to feature updated "Facts and Figures" along with the tools needed to both utilize data and contribute, including Informational Tools, Human Tools (list of experts, linking to SAA connect) and Software Tools.

CORDA will be meeting monthly via Zoom up to the annual meeting. The co-chairs are coordinating the work of the three subcommittees. We are awaiting notification regarding appointments of new members and the next co-chair.

Odum Institute Service Fees: https://odum.unc.edu/archive/



Service	Self-Service $0	Guided Service $3,000 + per dataset fee*	Lifecycle Service $5,000 + per dataset fee*
UNC Dataverse tool access	X	X	X
Data citation generation	X	X	X
Persistent identification (DOI)	X	X	X
Basic utilization reporting	X	X	X
Long-term preservation	X	X	X
Standardized metadata	X	X	X
User support	limited	standard	dedicated
Introductory Dataverse software training		X	X
Dataset collection arrangement		X	X
Metadata template development		X	X
Data Management Plan implementation			X
File format normalization			X
Data file review			X
Access policy enforcement			X
Education and training program development			X