Digital Scriptorium

Digital Scriptorium 2.0 Project Report I
November 6, 2020

Introduction

In July 2020, the IMLS awarded Penn Libraries a $100,000 National Leadership for Libraries Grant to oversee a one-year planning period for the future of Digital Scriptorium (DS), a consortium of US institutions who are dedicated to providing open access to their manuscript data and images. Since 1997, DS has hosted an online platform and database to facilitate this mission, but its purpose and technical infrastructure needs redevelopment in order to remain viable in the 21st century. The ultimate goal of DS is to become the national union catalog of pre-modern manuscripts in the US. Over the next year, the planning grant project team will refine the purpose and scope of the DS platform, develop a new data model for DS 2.0, and create plans for the technological and financial sustainability of the platform.

Three months into the planning period, we have established weekly technical and advisory meetings for the project, completed an institutional survey of current and potential DS member institutions, and hosted a series of virtual planning meetings attended by over 40 scholars, curators, librarians, and technologists from across the United States and Europe. We are on track for a successful completion of the planning grant in July 2021, when we will present a plan for DS 2.0 implementation to the DS Board of Directors.

Logistics of the project

The project team currently includes Lynn Ransom (PI), Emma Thomson (Project Manager), and Doug Emery (Technical Advisor). The day-to-day operations of DS 2.0 development are facilitated by two regular meetings. The DS 2.0 Steering Committee, a sub-committee of the

Digital Scriptorium Board of Directors, met weekly on Fridays from August through October 2020 to plan and produce the DS 2.0 Institutional Survey and Planning Meetings, discussed in more detail in a subsequent section of this report. Members of this group include DS 2.0 Project PI Lynn Ransom, Executive Director of DS Debra Cashion, DS 2.0 Project Manager Emma Thomson, DS Advisory Council Chair Lisa Fagin Davis, Ray Clemens (Yale), and Elizabeth Hebbard (Indiana U). After completion of the Planning Meetings in October, and reporting on these to the DS Board of Directors at their monthly meeting on October 16, the Steering Committee is transitioning to biweekly meetings, with a focus on continued advisory and discussion as the project moves towards its next tasks, the current DS data assessment and environmental scan.

A separate weekly meeting, dedicated to the technical nuts-and-bolts of the project, occurs every Wednesday. This meeting, attended mostly Penn Libraries staff, includes librarians, programmers, metadata specialists, and manuscript catalogers, who meet to discuss ideas for the technical infrastructure of DS 2.0, including potential technology stacks, data crosswalks, current standards (and the lack thereof) in manuscript cataloging practices across institutions, and more. In the upcoming months, this group will devote its attention to the workflow for DS 2.0's data ingestion, in advance of the data modelling work which will begin in January 2021.

Institutional Survey

DS invited all institutions in the United States with pre-modern manuscript holdings to complete a survey on their needs for a union catalog of pre-modern manuscripts. The survey gathered metrics on the scope of manuscript collections in the US and the technologies used to describe and digitize them. We invited any US institution with pre-modern manuscripts to complete this survey, regardless of the size or scope of their collection. The survey was produced using a Google Form and publicized on the DCRM-L listserv and the DS internal mailing list. Responses were collected over a three-week period, from September 8th - 30th, 2020. 35 institutions completed the survey, with roughly half of them being current DS members or associates. Survey responses were presented by Emma Thomson during the first day of the DS 2.0 Planning Meetings. Graphs of the survey results are included in Appendix A.

We organized the survey into three sections based on the subject matter of the questions each section contained. The first section included questions about the current state of each institution's DS membership, and what benefits and challenges institutions faced in initiating or maintaining DS membership. [1] Respondents consistently mentioned the increased accessibility, visibility, and discoverability of their collections as a key benefit of DS membership. Financial concerns emerged as the main challenge to DS membership, with limited staff resources to facilitate DS data creation and maintenance as another common theme.

---

[1] See Appendix A, p. 8-9, for survey results related to this section.

The second survey section gathered metrics on the size, scope, and cataloging practices of the respondents. The majority of institutions have a total materials budgets of $8 million or less, with a wide range of student body populations, ranging from no students at all to over 30,000.[2] Starker contrasts emerged regarding the size of the manuscript collections: 37% own over 600 manuscript objects, with the next most common response being between 5-50 manuscripts (26% of respondents).[3] The regional scope of their collections indicated a strong emphasis on European material. Every institution held at least one item from Europe, with that region making up 81-99% of the collection of 22 respondents.[4] But 24 institutions also own at least one item from the Middle East, and over 10 institutions hold manuscripts produced in Africa, the Americas, and Asia.[5] 25 institutions (71% of respondents) already have an existing online catalog for their manuscripts, either in their institution's general OPAC or a specialized catalog dedicated just to their manuscripts.[6] Yet only 63% of respondents report that the majority of their collections are cataloged, indicating that for many manuscripts in the US, no data about them exists yet.[7] The percentage of manuscripts currently digitized at each institution were even more varied: 34% of institutions reported that less than 25% of their manuscript collections were digitized, but the next most common response, at 23% of respondents, stated that between 75-100% of their collections were digitized.[8] Despite this range in manuscript digitization, a majority of respondents (77%) stated that their institutions have the equipment and personnel in place for digitization, indicating that manuscripts are not a priority for digitization at many institutions, or their state of conservation does not permit them to be digitized.[9]

The survey questions related to cataloging practices deserve close attention, as the technical infrastructure of DS 2.0 will in many ways depend on the standards and practices of the data it hopes to ingest. Of the myriad data formats available these days, MARC is still overwhelmingly the most common standard used to describe manuscript objects among our survey respondents (27 out of 35 total institutions). The next most common response (from 17 institutions) was unstructured text documents including PDF or Microsoft Word files. Zero respondents currently publish descriptions in BIBFRAME or RDF, a telling sign of the lack of linked data currently available about manuscripts.[10] Institutions did report widespread use of controlled vocabularies in their manuscript descriptions, including the Library of Congress Subject Headings and Genre Form Terms, the Getty Vocabularies, VIAF, and the RBMS Vocabularies. DS 2.0 should aim to align with these controlled vocabularies as much as possible to facilitate easy data migration and discoverability. We also asked survey respondents about their use of IIIF, with a little more than half of institutions (51% and 54%) reporting that they

---

[2] Figures 5 and 6, p. 10.
[3] Figure 7, p. 11.
[4] Figure 12, p. 13.
[5] Figures 8, 9, 10, and 13, p. 11-14.
[6] Figure 14, p. 14.
[7] Figure 16, p. 15.
[8] Figure 17, p. 16.
[9] Figure 18, p 16.
[10] Figures 19, 20, and 22, p. 17-18.

currently have the ability to store and publish IIIF-compliant images, respectively.[11] Only 23% of respondents reported that their digital repositories are OAI-PMH compliant, which will limit the ability of DS 2.0 to harvest data according to that protocol.[12] Happily, over 80% of respondents indicated that their institutions were willing to make their manuscript data and images open access, with 19 of those institutions willing to support no restrictions on their content at all.[13]

The final survey section included questions about the functions and features institutions hoped to see in DS 2.0. We asked respondents what level of metadata would be sufficient for a DS 2.0 manuscript record, and provided 3 response options: full description, brief identifying description with links, or other (fill-in-the-blank). 70% of respondents chose the full description option, with 3 institutions contributing a fill-in-the-blank response (two of these other responses were amicable to a brief record).[14] When asked about the level of digitization that would be sufficient for a DS 2.0 manuscript record, 40% of respondents stated that a sample selection of images would suffice, with 29% preferring cover-to-cover digitization, and 23% responding that no digitization is necessary.[15] We also asked institutions to rank the importance of DS 2.0's ability to collaborate with international partners and projects, as well as utilize linked open data technologies. The ranking scale ranged from 1-5, with 1 being not important and 5 being extremely important. The majority of respondents ranked both of these questions at a 4 or 5, indicating these issues are very important.[16] We ended the survey with a fill-in-the-blank question, asking respondents what other features and functions they would like to see in DS 2.0. We received a wide range of responses here, including a lot of "I don't know" answers. The most common response was the need for authority control (6 answers), with other comments including developing easier ways to contribute to DS, linking DS records to other projects, and better documentation for new users/contributors.[17]

The survey responses, overall, reveal the broad range of institutional practices that DS 2.0 will need to accommodate in its workflow and data model development. Still, some key agreements among the responses are visible. The dominance of MARC as a data format necessitates that DS 2.0 design a way to easily get data out of MARC records. Complicating this issue is the fact that although nearly every institution uses MARC, there can be wide variability in how each institution implements specific fields in their own records. For example, call number and location information may appear in a number of different fields across MARC records at different institutions. Even within institutions, standards change over time, and MARC records created ten years ago may follow different standards than records created recently. Another important lesson from the survey is the lack of linked open data currently being created about manuscripts. This is both a problem and an opportunity for DS 2.0. Widespread use of linked data would make it much easier to build the new platform. However, the lack of linked data

---

[11] Figures 25 and 26, p. 20.
[12] Figure 27, p. 21.
[13] Figures 28 and 29, p. 21-22.
[14] Figures 30 and 31, p. 22-23.
[15] Figure 32, p. 23.
[16] Figures 33 and 34, p. 24.
[17] Figure 35, p. 25.

affords DS 2.0 a great opportunity to become the central linked open data provider about manuscripts in the US, in addition to becoming the national union catalog. Institutions and researchers of all types would benefit from this, as linked open data offers superior interoperability and querying potential to traditional relational databases. Survey respondents overwhelming preferred that DS 2.0 manuscript records include full manuscript descriptions, rather than only a brief amount of identifying information. In the Planning Meetings, we generally received the opposite response from participants, who agreed that managing full scholarly descriptions of manuscripts is neither warranted nor possible for DS 2.0. The Planning Meetings, discussed in detail in the following section, afforded participants four days to think deeply about the opportunities and challenges the next iteration of DS faces, which in all probability accounts for the gulf we see between the opinion of the survey respondents and the participants of the Planning Meetings.

Planning Meetings

DS invited its members and other stakeholders to gather together for a series of virtual planning meetings about DS 2.0 on October 6, 8, 13, and 15 from 1-3pm ET. These meetings were included in the initial grant proposal and intended to be held locally in Philadelphia, but the COVID-19 pandemic forced them to be held remotely instead. The first three days featured presentations from experts about related projects in the field, as well as breakout group work where attendees worked in small groups to discuss the central questions of the meetings. The final meeting day was dedicated to general discussion about the future of DS, with each of the four breakout groups presenting a summary of their conversations over the past three days. 63 people registered for these meetings, with consistent turnout of around 40 people each day. Surprisingly, the day with the highest turnout was the fourth and final day of the meeting, with 43 people in attendance. The sizeable and consistent attendance at these meetings demonstrates a vested interest in the DS 2.0 project among the manuscript studies community. The Appendix of this report includes a copy of the Planning Meeting Agenda and the Breakout Group Activity handout, which lists all of the different questions we asked the breakout groups to discuss together.[18]

While these meetings were chiefly designed to include the wider manuscript community in the project development process, they also served as a chance to manage that community's expectations for what DS 2.0 can achieve. We asked the breakout groups to brainstorm about the future of DS without restriction ("blue-sky thinking"), but only after they attended presentations about real-world projects, all of which encounter certain conceptual, organizational, and technical limitations. We wanted the groups to consider all of the possibilities for what a national union catalog can and should be, while at the same time discussing the challenges that it will certainly face.

---

[18] See Appendix B, p. 26-27, and Appendix C, p. 28-29, respectively.

Each breakout group presented a summary of their discussions on the final day of the planning meetings. We provided a long list of topics for their discussions, and no group was able to discuss or find solutions to every question we provided.[19] Many points of general consensus appeared among all of the groups: that the scope of DS 2.0 should be global, rather than focused only Western European material; that DS should facilitate more services to support institutions with little or no manuscript staff or cataloging expertise; that DS should increase its visibility and role as a community of institutions and people who promote open access to manuscript data. The groups offered ideas about what a core DS 2.0 manuscript record should contain: one group suggested a middle-way approach between a full manuscript description and a brief identifying record; another group advised that the only mandatory fields should be a unique identifier and the location of the manuscript. Two of the groups argued for a more precise articulation of the purpose of DS 2.0, with one group arguing that the platform's main goal should be to help users find manuscripts (with any deeper research or data analysis a secondary concern only), with a separate group echoing that idea by proposing that DS 2.0 should be the start of a user's research process, not a one-stop-shop for official manuscript descriptions and cover-to-cover digitization.

After the breakout group presentations, we ended the meetings with a general conversation about the future of DS and the way forward as we develop DS 2.0. Most attendees agreed that the main goal of DS 2.0 should be comprehensive coverage of the manuscript holdings in the United States. This means that the focus will be on a minimal set of required metadata elements for each record, simultaneously lowering the bar for participation (since full scholarly catalog records for each manuscript won't be required), while also reducing the overhead required for the development and maintenance of the dataset. DS 2.0 can also become more flexible and efficient by utilizing IIIF compliant images produced by institutions, rather than storing and managing images directly in the DS platform. Participants agreed that DS 2.0 cannot be everything to all users, but instead must focus on providing what is most needed out of a national union catalog at this time: standard identifiers for manuscripts and information about where they are located in the United States.

Conclusion and Looking Ahead

The project team has successfully completed all scheduled tasks as planned during the first three months of the grant, including the institutional survey, the planning meetings, and the establishment of regular meetings to facilitate ongoing brainstorming conversations. Over the past few months, the project team's understanding of the core issues of the planning period has changed. At the outset of the project, the development of a new data model for the DS platform seemed to be the central problem that would need to be tackled over the coming year. We now understand that the development of an efficient and repeatable workflow for ingesting data into the DS platform is the more acute issue. The next three months of the work schedule will also be dedicated to completing an assessment of the current data in DS and an

---

[19] Appendix C, p. 28-29.

environmental scan of related union cataloging projects that have been completed or are in progress. This research will give the project team a precise understanding of the limitations and challenges of the current DS data, and how other related projects have tackled similar issues to what we are facing regarding the development of a data ingestion workflow and data model. The second DS 2.0 Project Report will be published in February 2021.
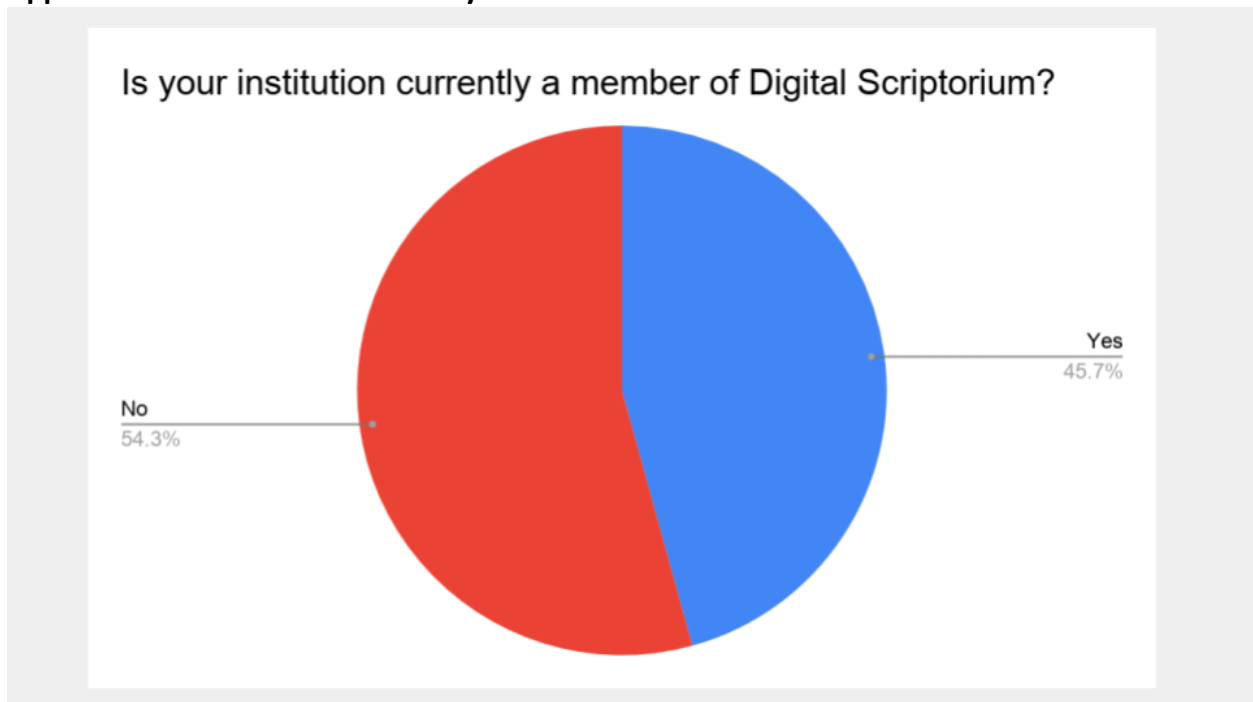
Appendix A. DS Institutional Survey Results

Figure 1. Current state of DS membership
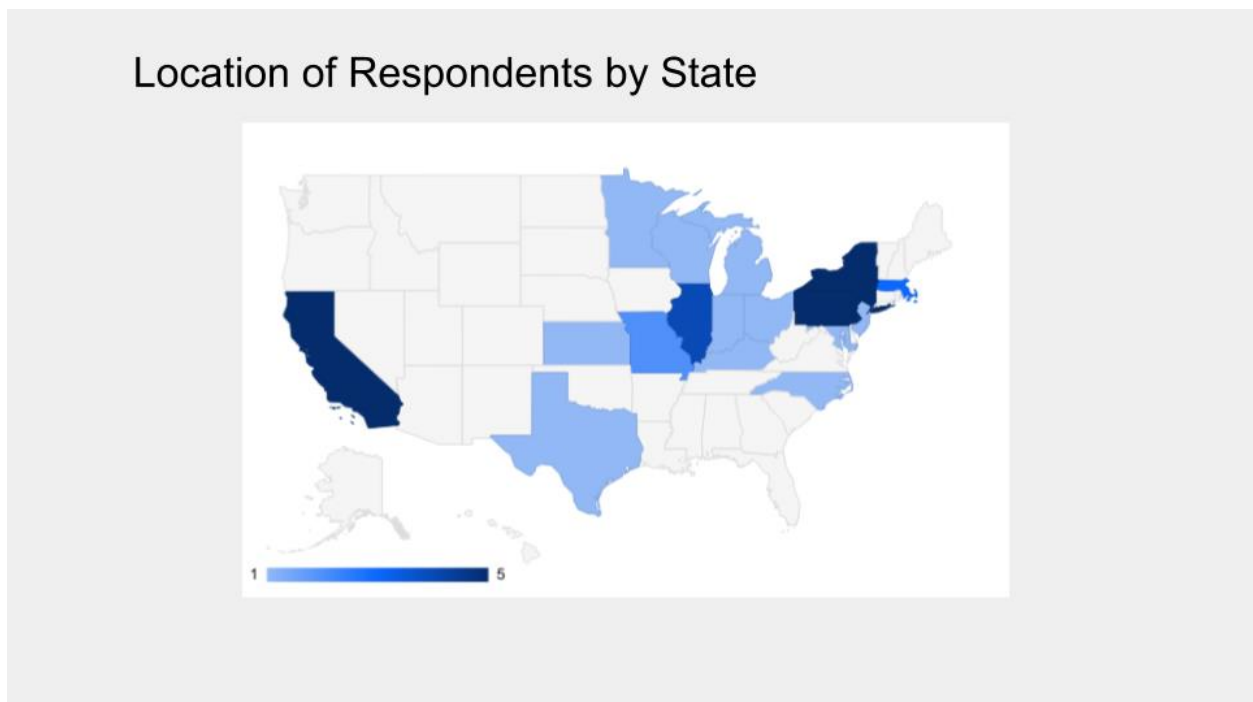
Figure 2: Location of Respondents by State

Figure 3: Benefits of maintaining a DS membership

Figure 4: Challenges of maintaining a DS membership
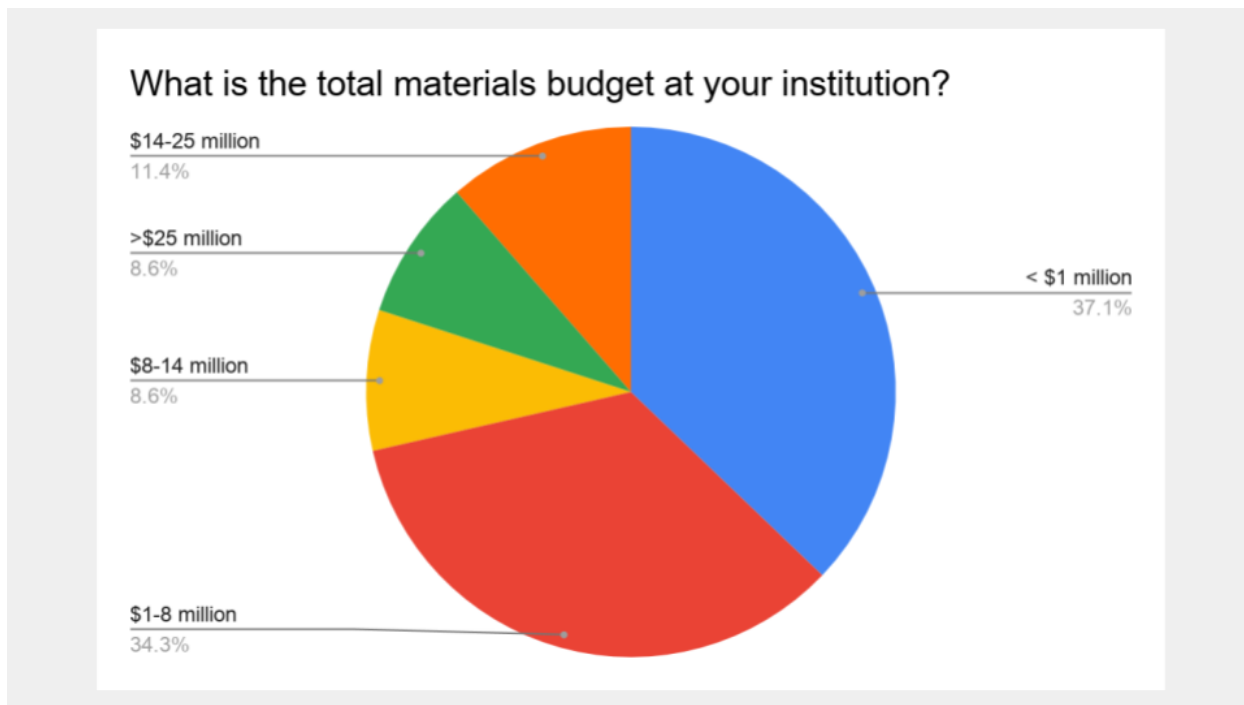
Figure 5: Total materials budgets
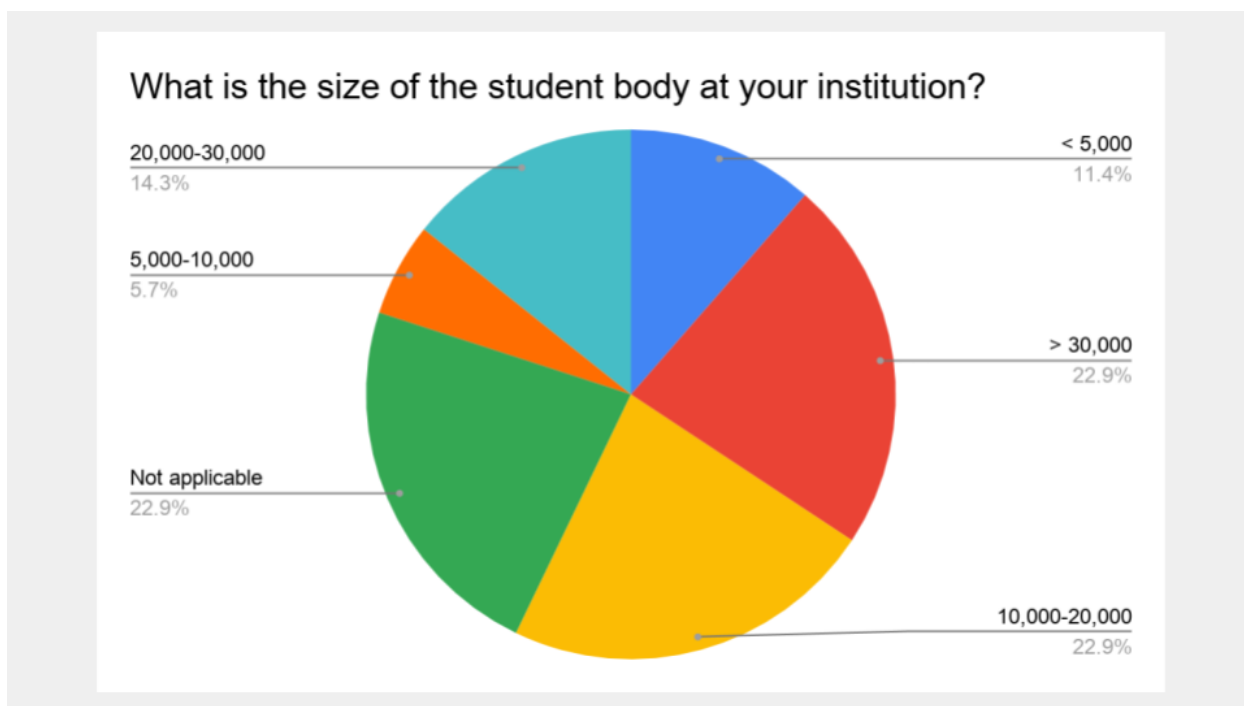
Figure 6: Student body size
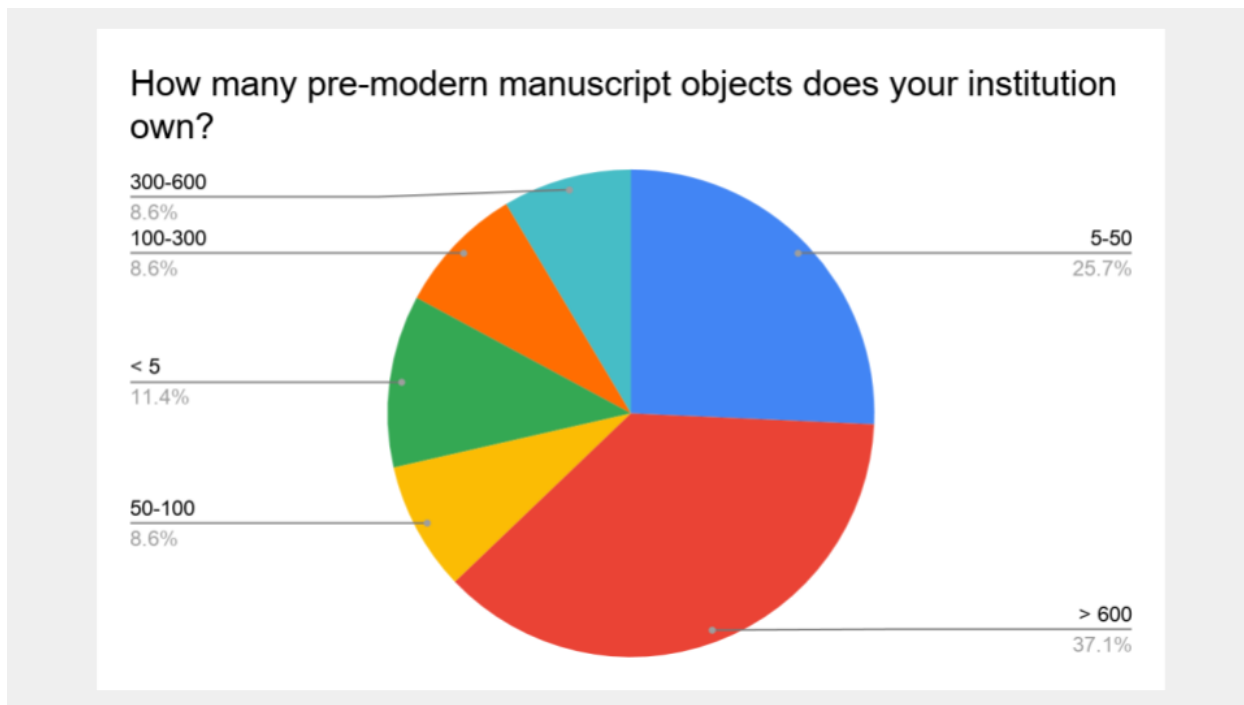
Figure 7: Size of pre-modern manuscript collections
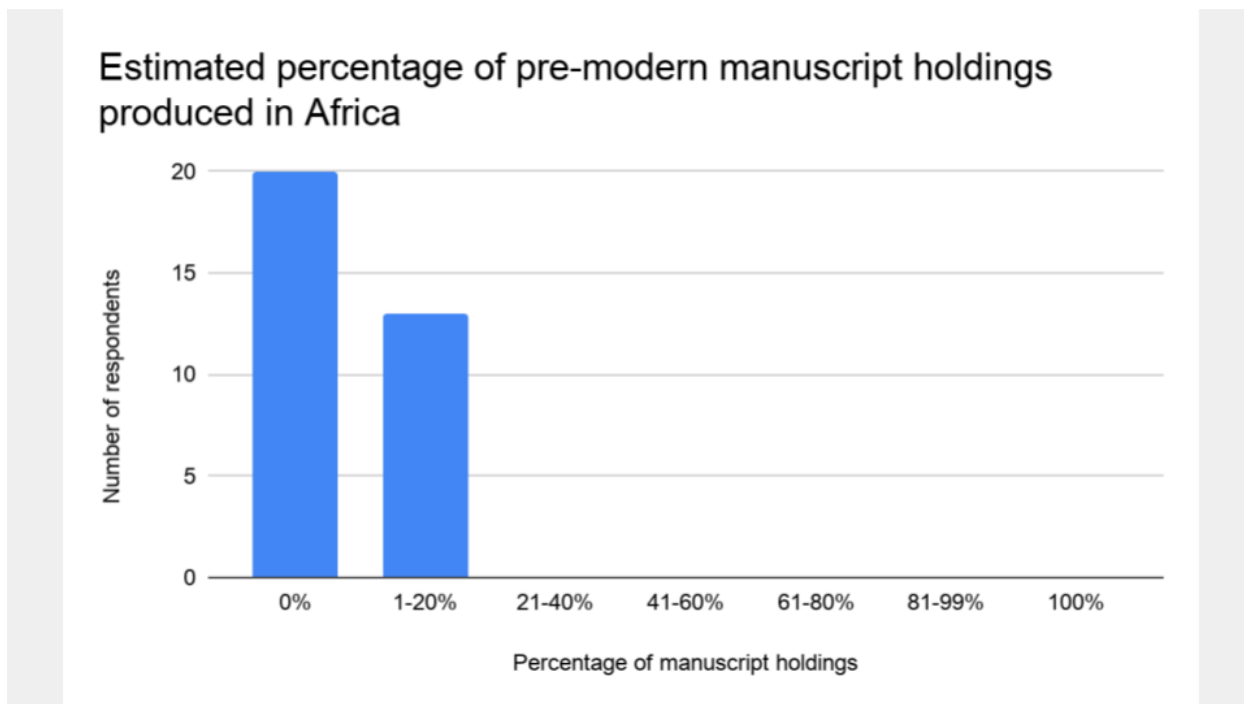
Figure 8: Percentage of pre-modern manuscript holding produced in Africa

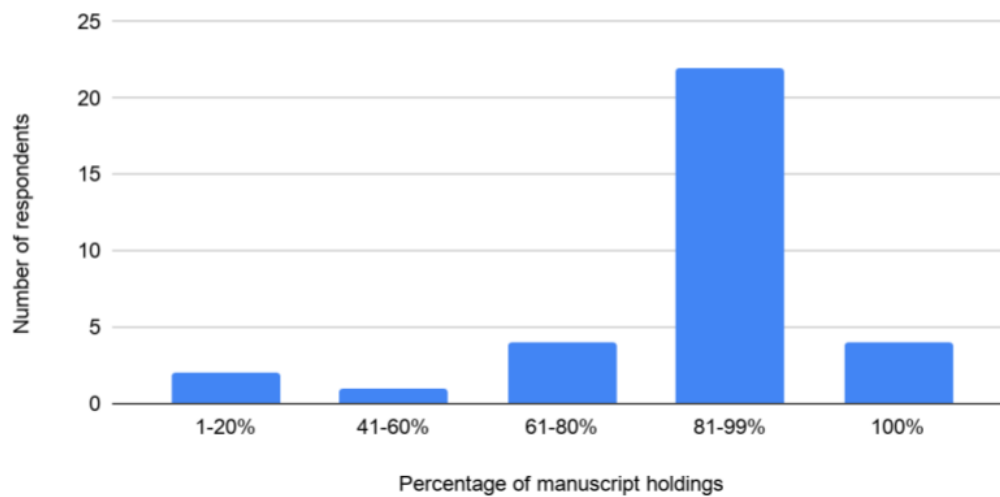**Estimated percentage of pre-modern manuscript holdings produced in America**
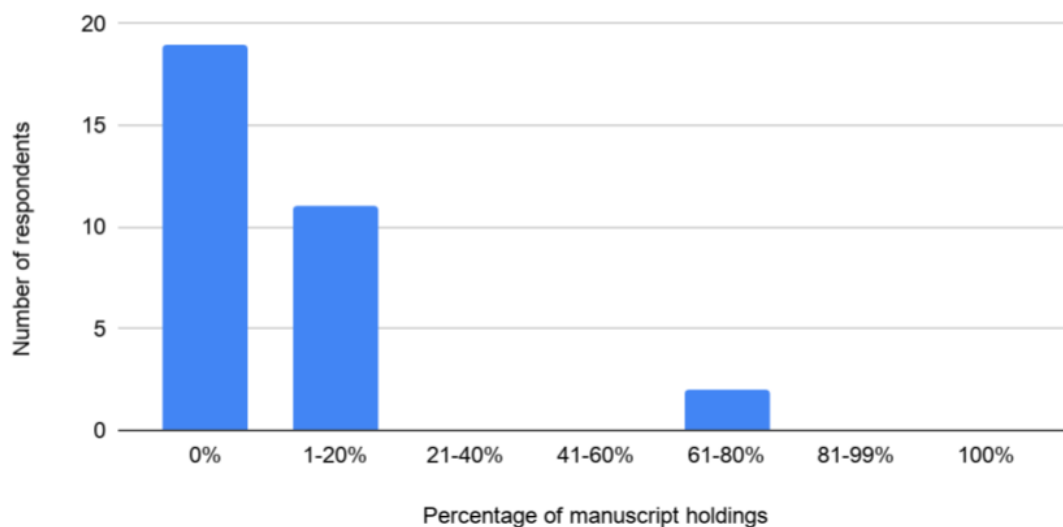
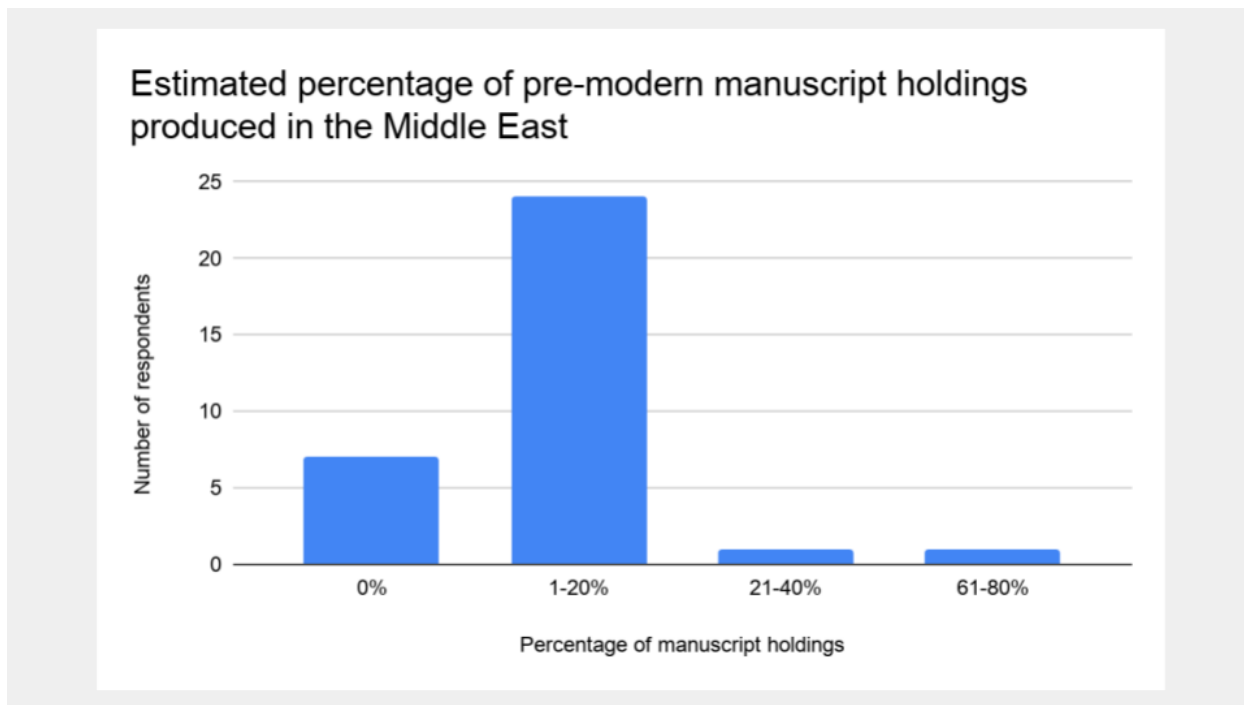Figure 9: Percentage of pre-modern manuscript holdings produced in the Americas

Estimated percentage of pre-modern manuscript holdings produced in Asia
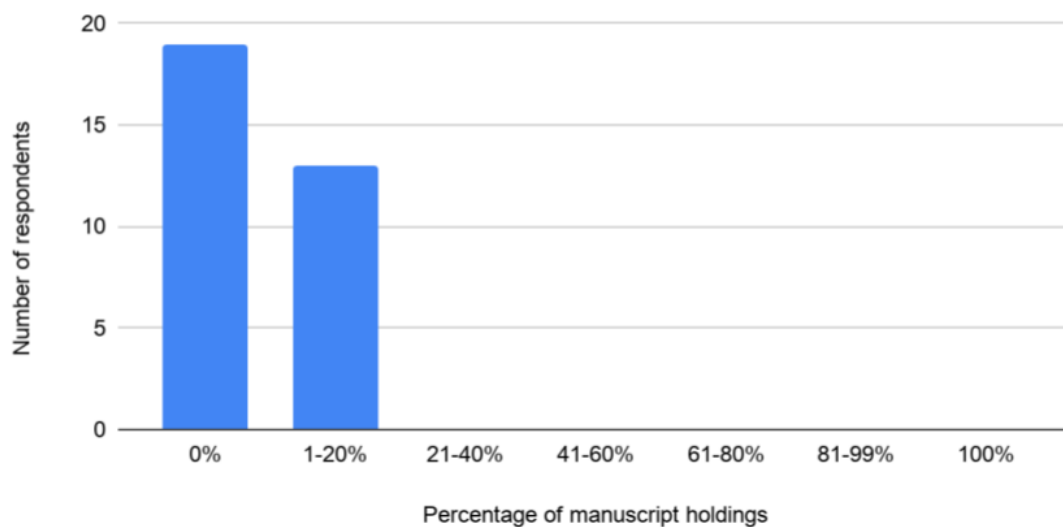
Figure 10: Percentage of pre-modern manuscript holding produced in Asia

Estimated percentage of pre-modern manuscript holdings produced in Australasia
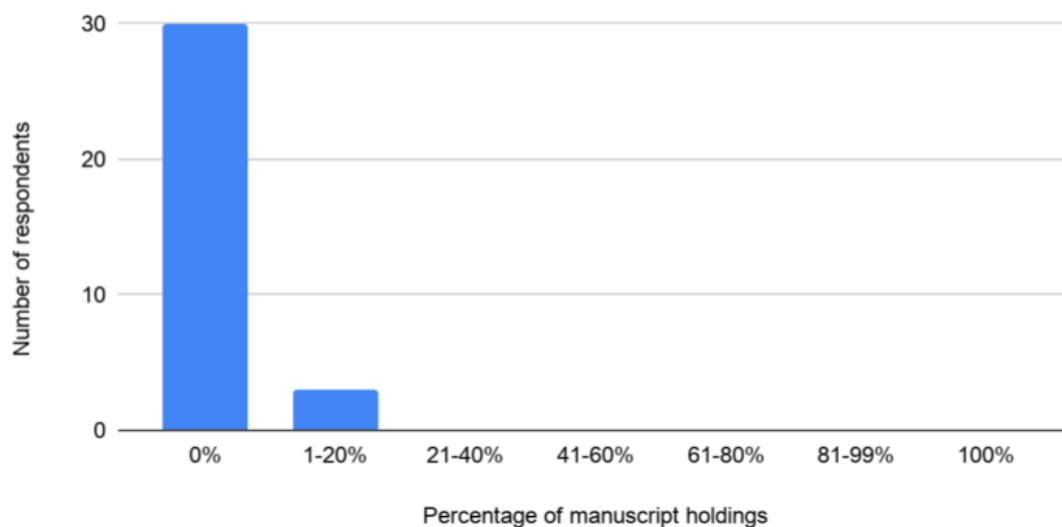
Figure 11: Percentage of pre-modern manuscript holdings produced in Australasia

Estimated percentage of pre-modern manuscript holdings produced in Europe

Figure 12: Percentage of pre-modern manuscript holdings produced in Europe

Figure 13: Percentage of pre-modern manuscript holdings produced in the Middle East
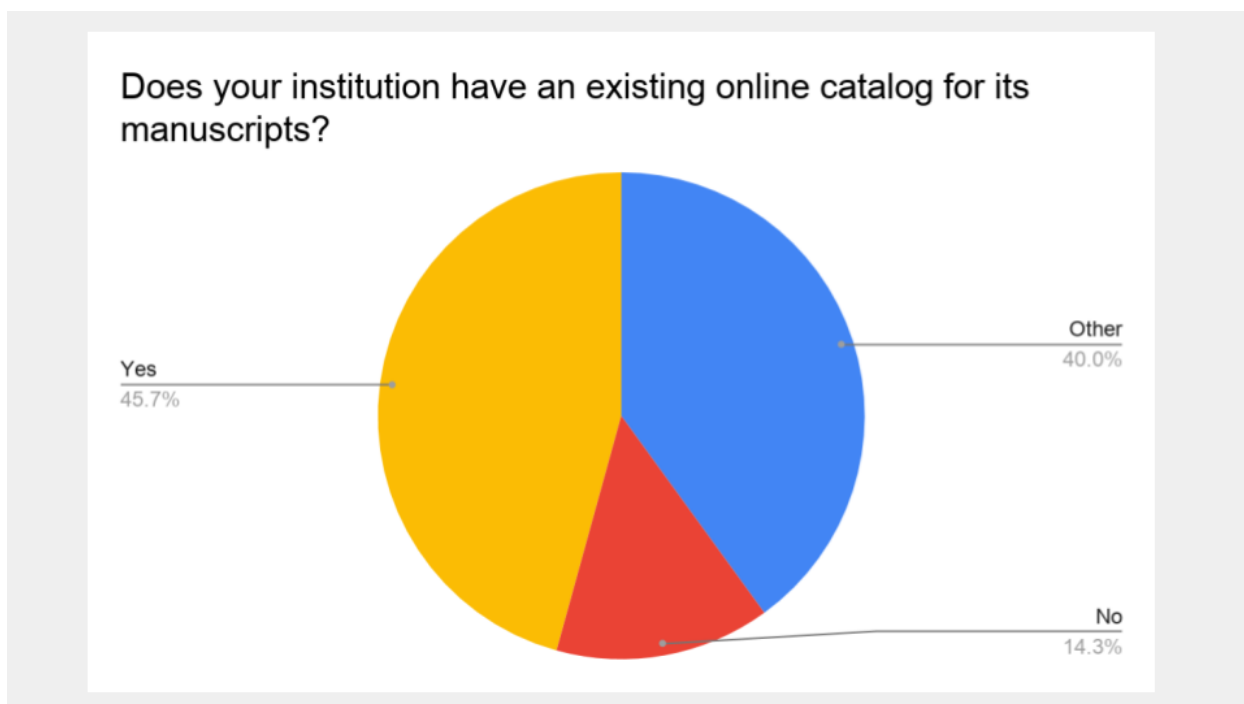

Figure 14: Existing institutional catalogs for manuscripts

Figure 15: Open-ended responses to presence of institutional catalogs for manuscripts
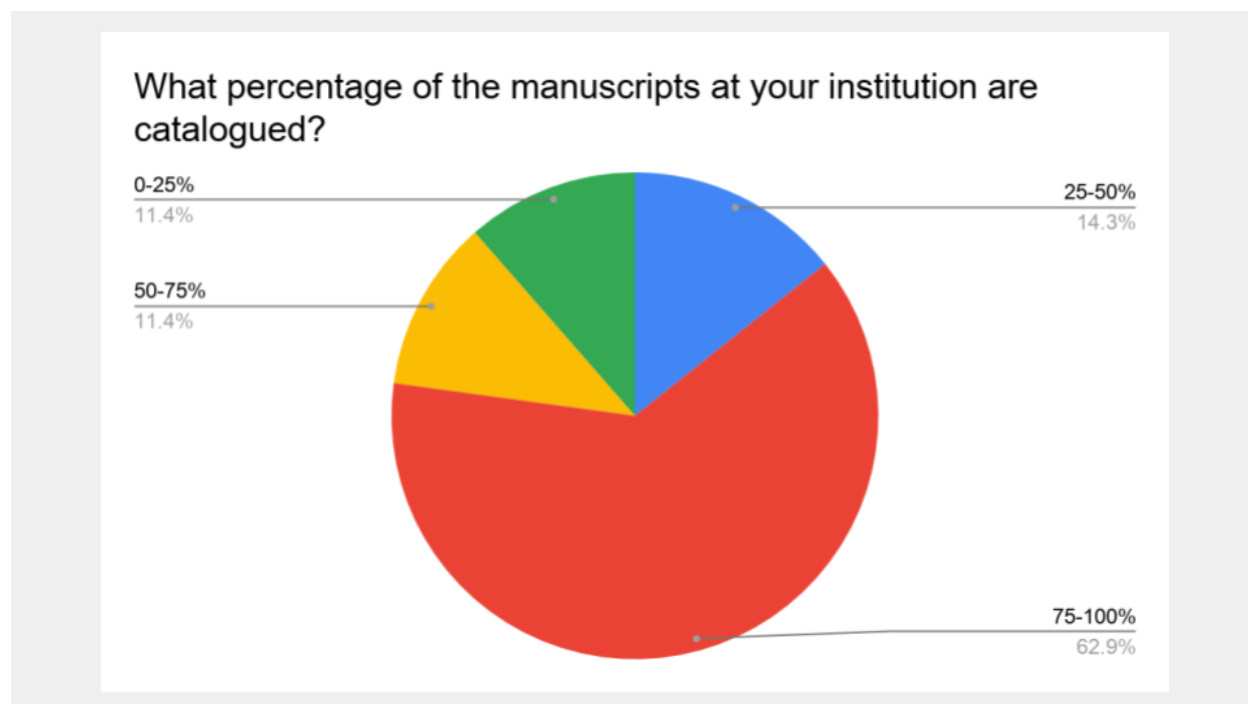
Figure 16: Percentage of manuscripts that are catalogued at responding institutions
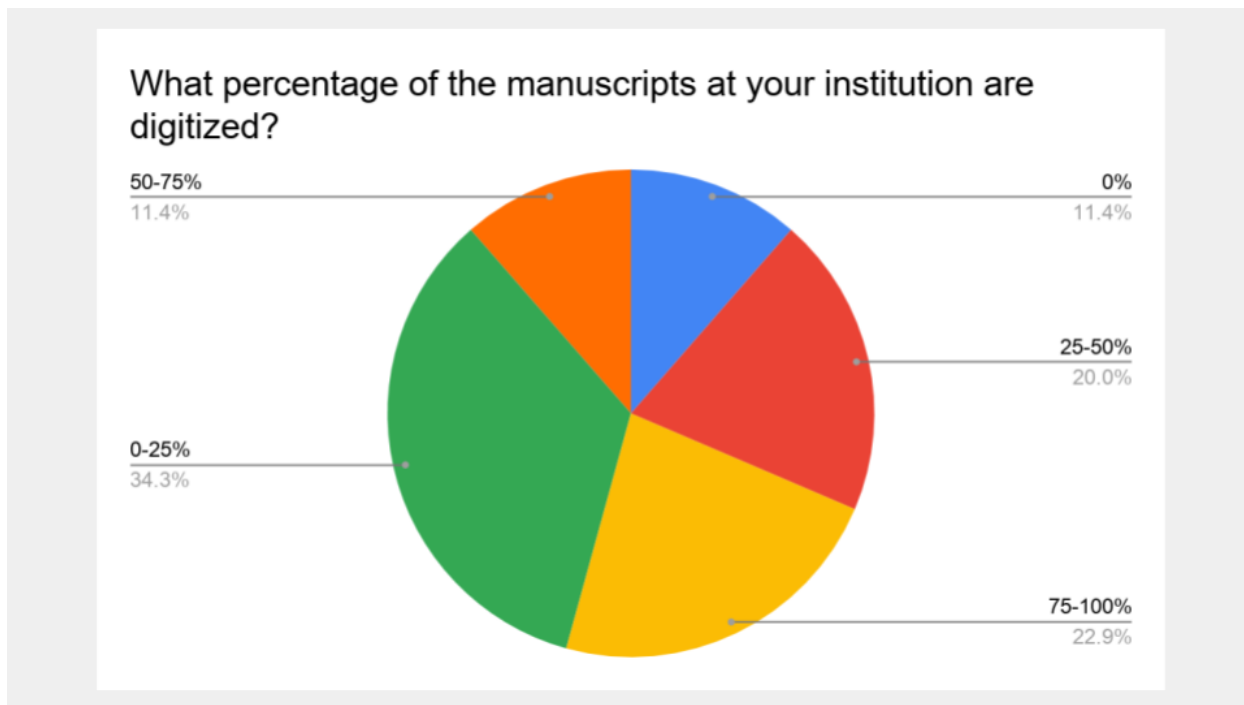
What percentage of the manuscripts at your institution are digitized?

50-75%
11.4%

0%
11.4%

25-50%
20.0%

0-25%
34.3%

75-100%
22.9%

Figure 17: Percentage of manuscripts that are digitized at responding institutions

Does you institution have equipment and personnel in place for the digitization of special collections?

Uncertain
5.7%

No
17.1%

Yes
77.1%
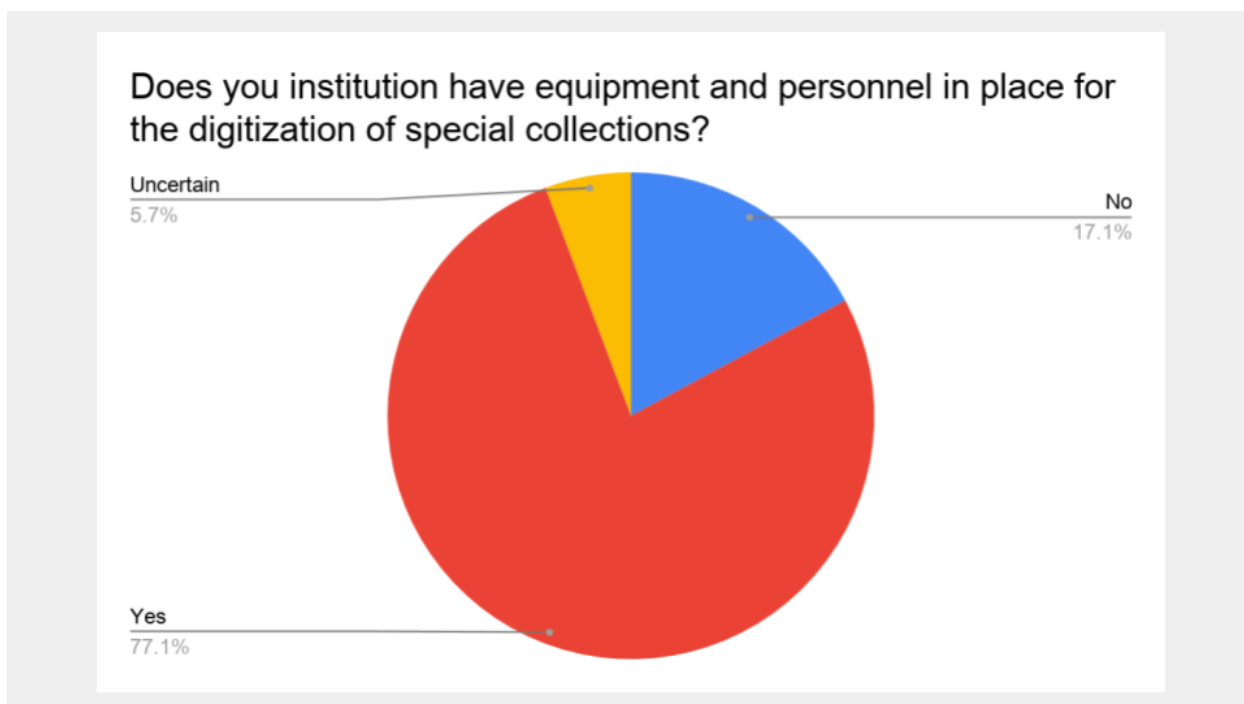
Figure 18: Digitization equipment and personnel in place at institutions

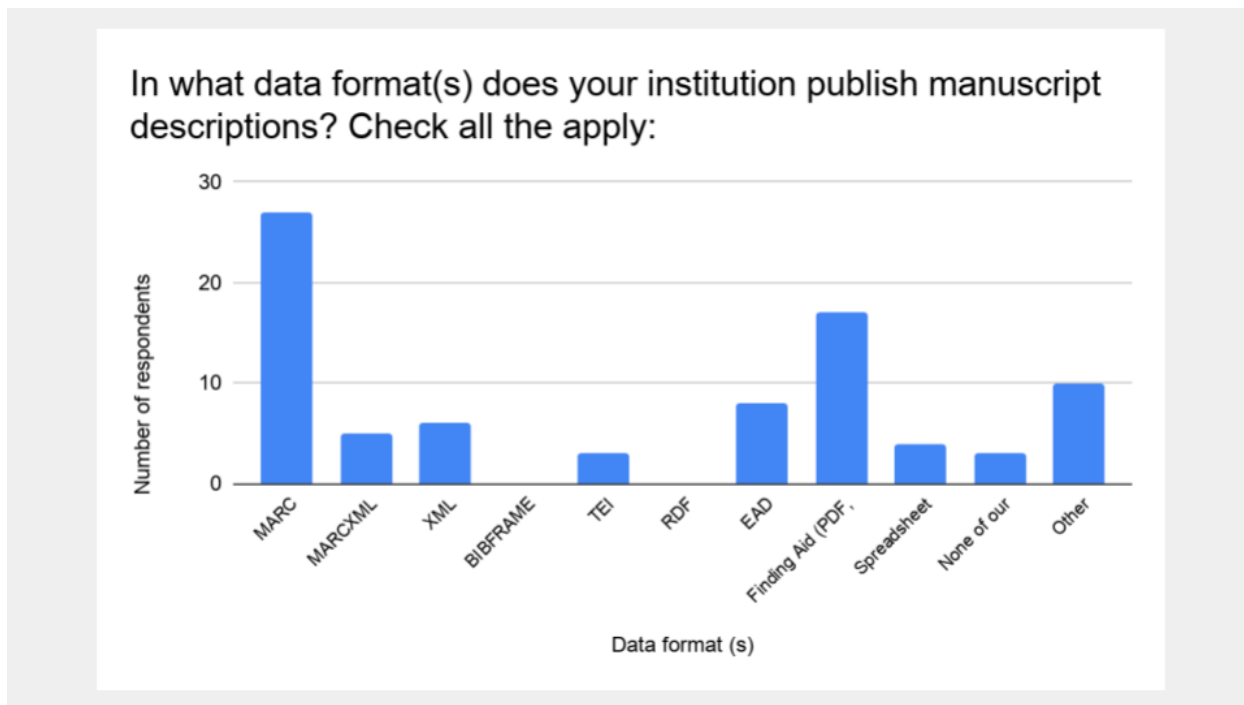In what data format(s) does your institution publish manuscript descriptions? Check all the apply:

Figure 19: Data formats used to publish manuscript descriptions (respondents could check as many options as were applicable)

Summary of other responses: In what data format(s) does your institution publish manuscript descriptions?

- MODS (3 responses)
- ContentDM (1 responses)
- DACS (1 response)
- JSON (1 response)

Figure 20: Summary of other data formats used to publish manuscript descriptions
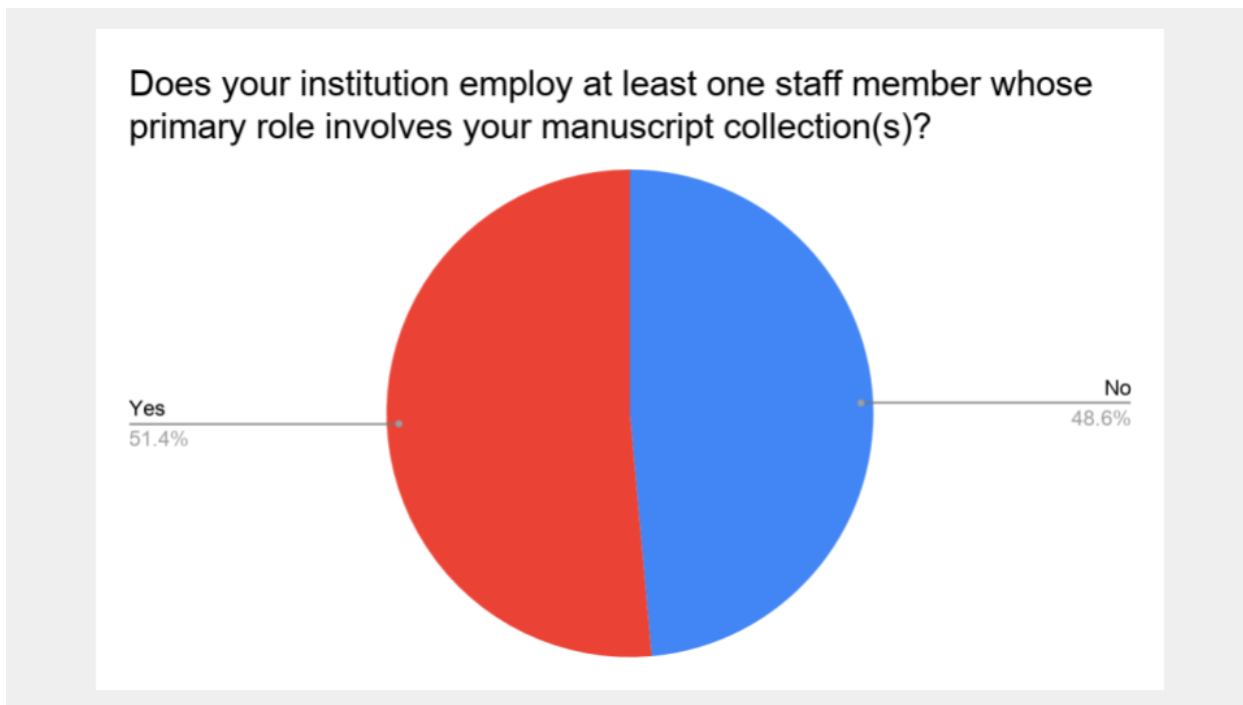
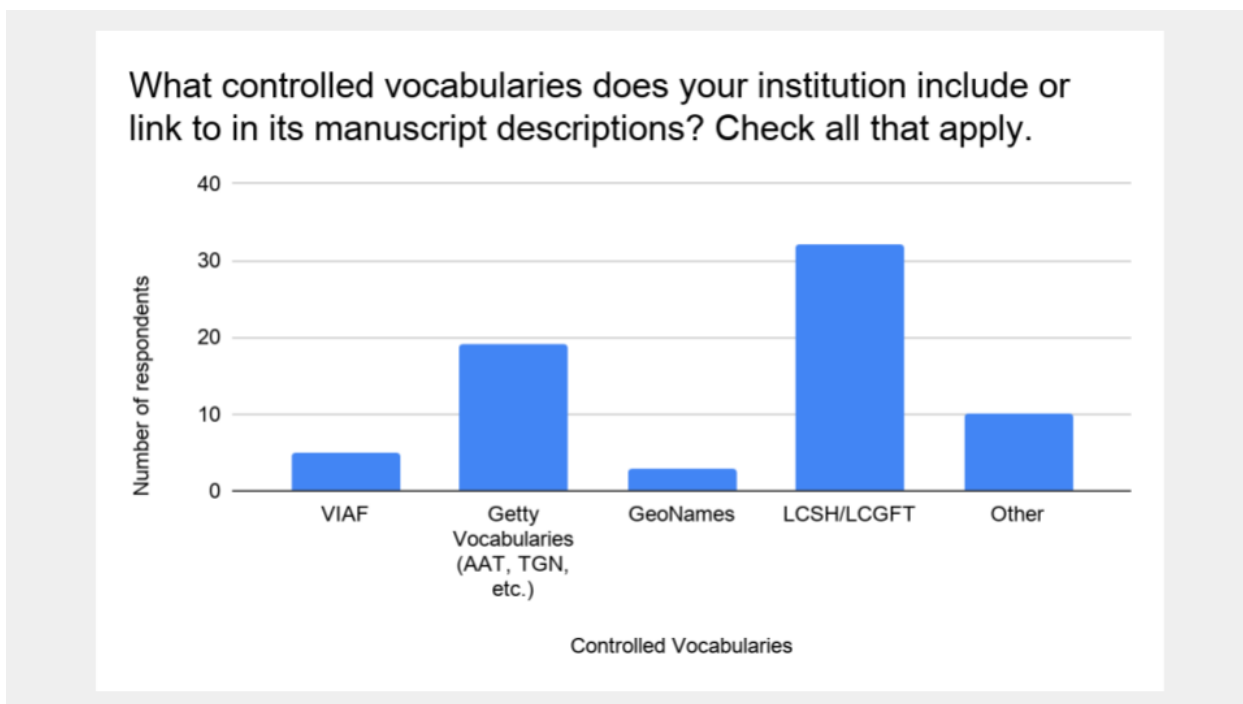Figure 21: Percentage of institutions who employ at least one staff member dedicated to manuscript collection

Figure 22: Controlled vocabularies used in manuscript descriptions

Figure 23: Other controlled vocabularies used in manuscript descriptions
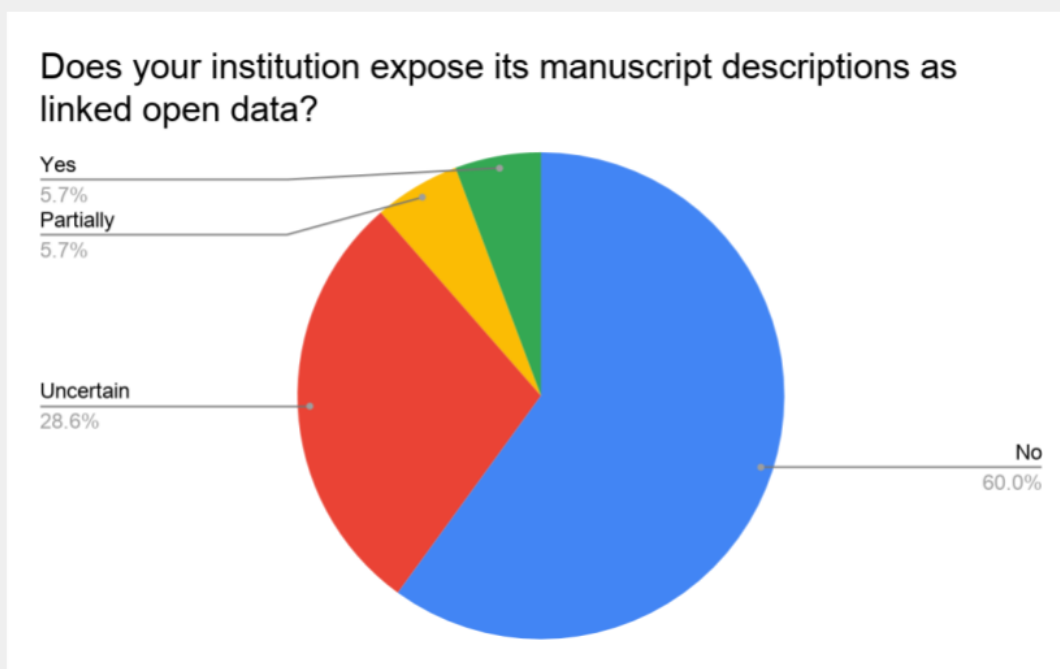
Figure 24: Institutional use of linked open data for manuscript descriptions

Does your institution have the ability to store its own IIIF-compliant images?

No
14.3%

Uncertain
34.3%

Yes
51.4%
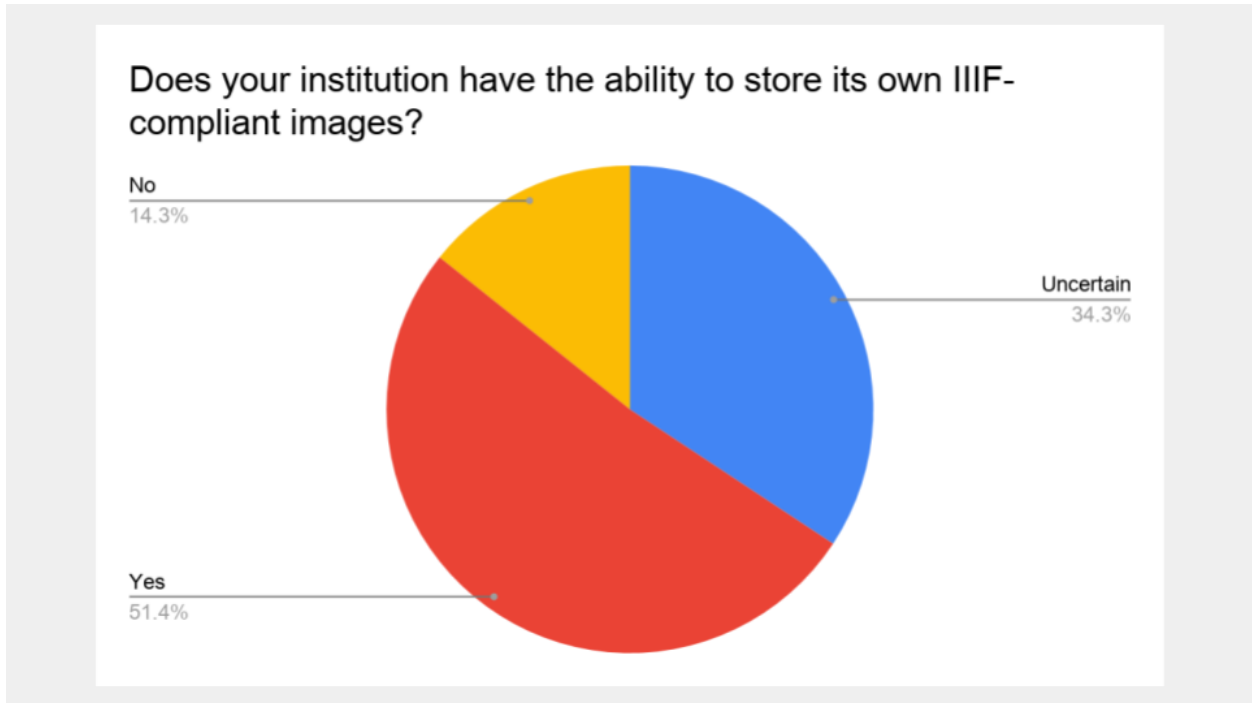
Figure 25: Ability of institutions to store their own IIIF-compliant images

Does your institution have the ability to publish its own IIIF-compliant images?
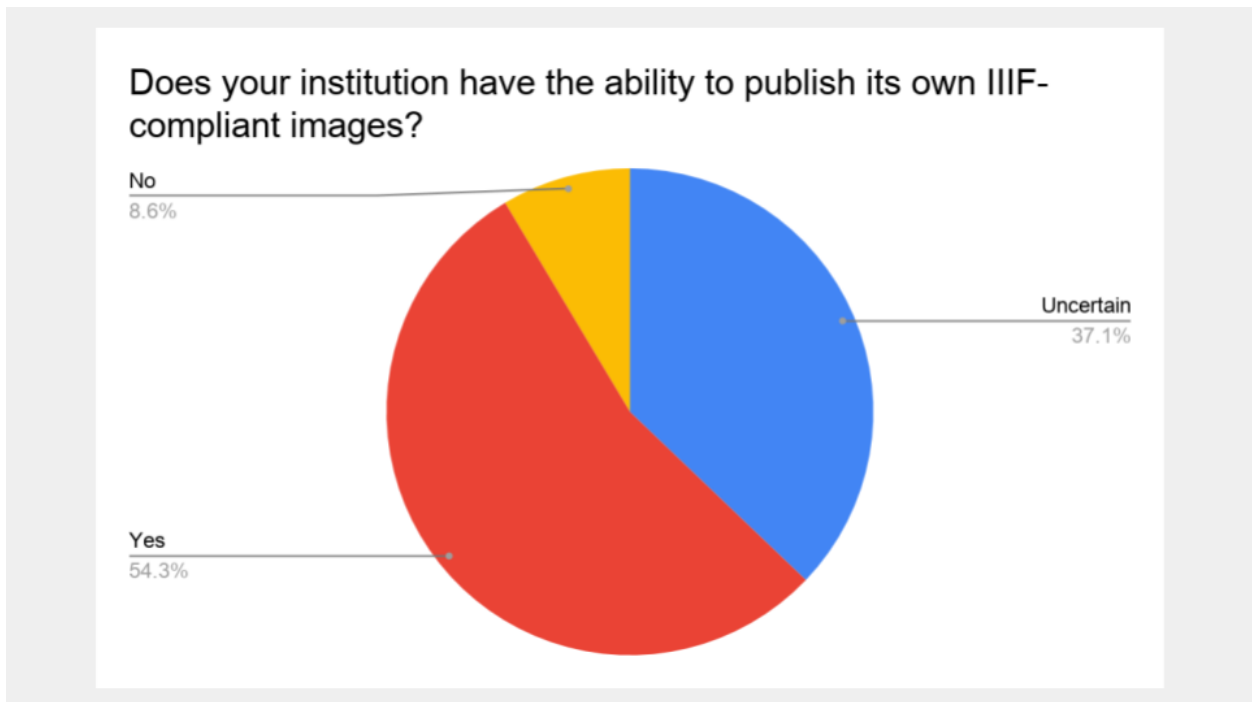
No
8.6%

Uncertain
37.1%

Yes
54.3%

Figure 26: Ability of institutions to publish their own IIIF-compliant images
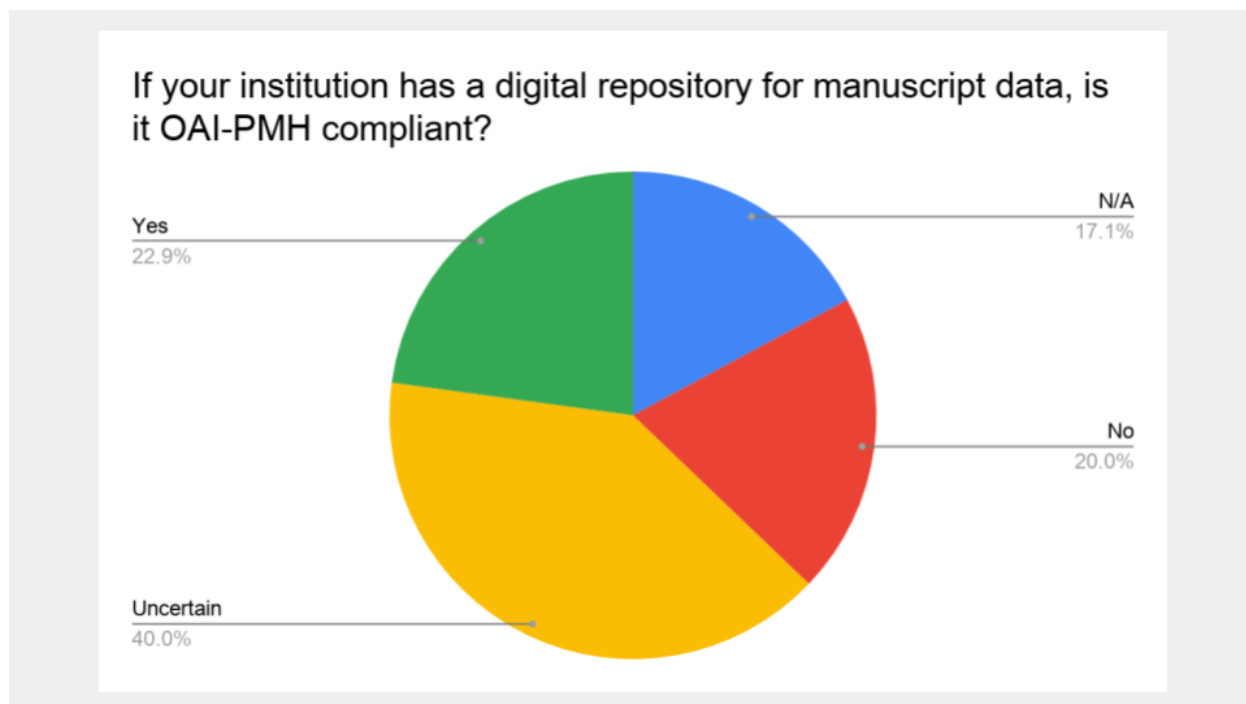
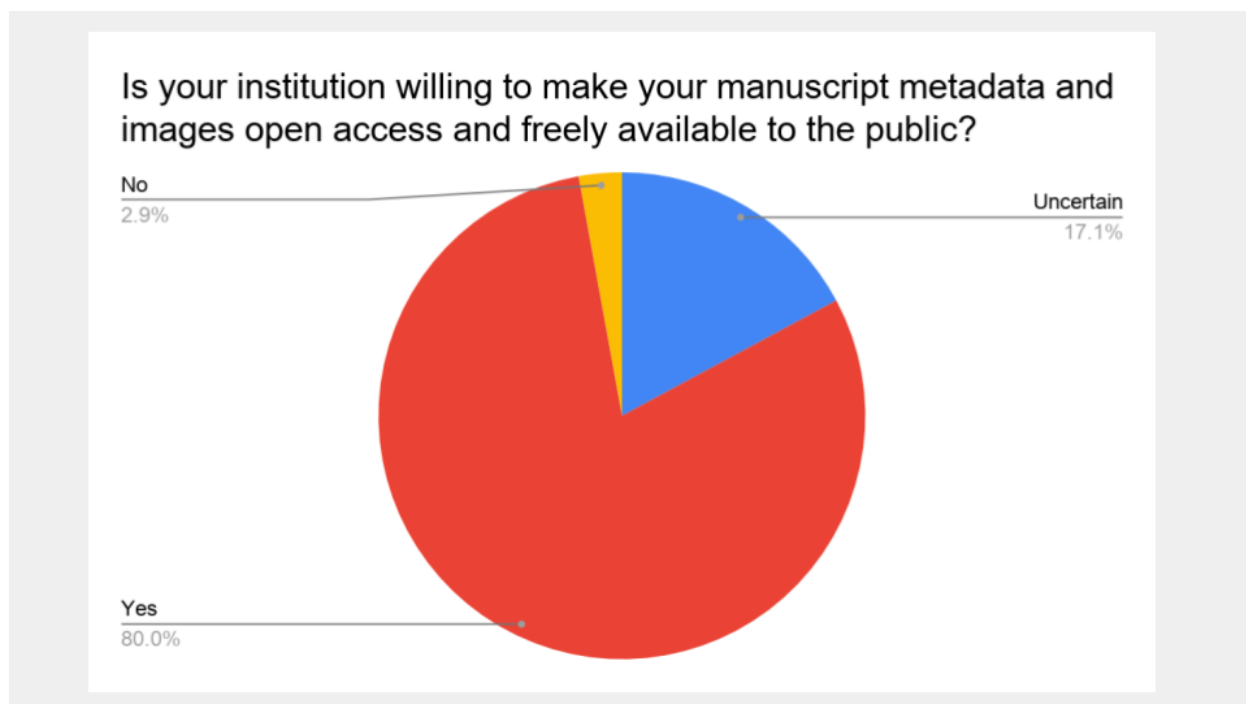Figure 27: Use of OAI-PMH standard among US institutions

Figure 28: Willingness of institutions to make their manuscript data and images open access

Summary of responses: "Please explain the level of open access your institution would be willing to support (for example, all metadata and images available for download and personal use, metadata and images available but with restrictions, etc.)"

- No restrictions (19 responses)
- Some restrictions (9 responses)
- Uncertain (7 responses)

Figure 29: Summary of open-ended responses to the level of open access institutions are willing to support

What level of metadata would be sufficient for a DS 2.0 manuscript entry for your collections?

Other
8.3%

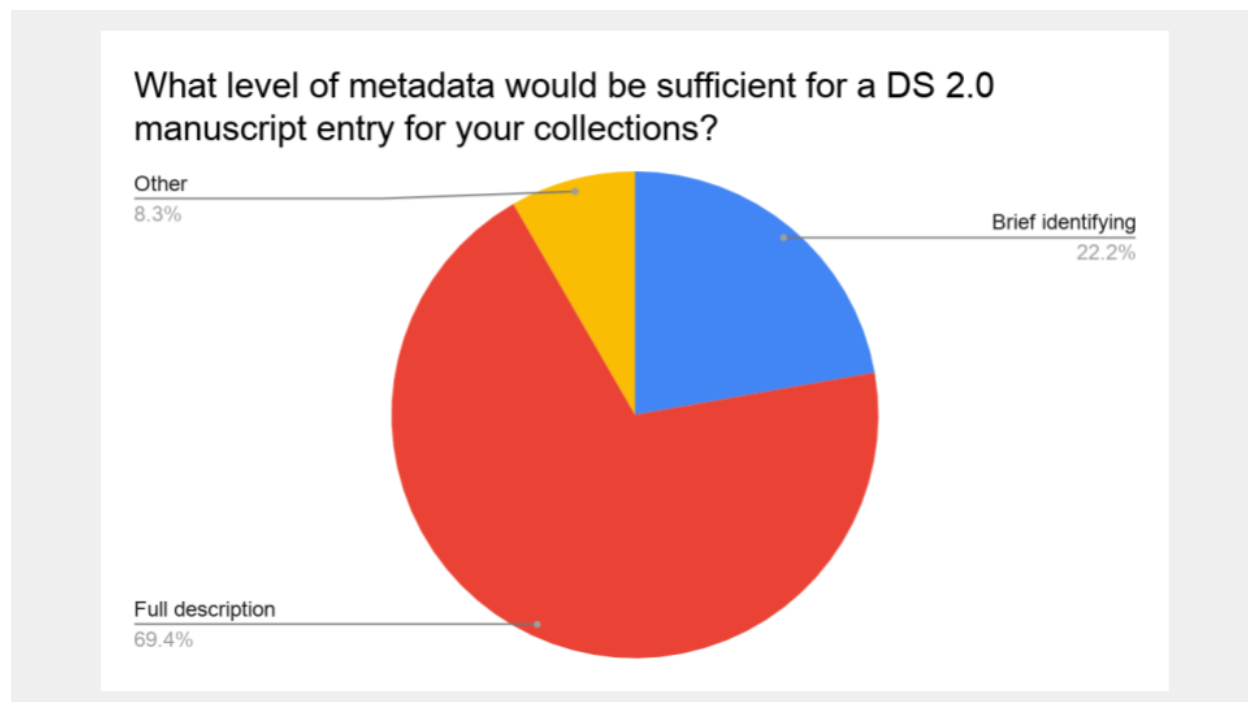Brief identifying
22.2%

Full description
69.4%

Figure 30: Level of metadata that would be sufficient for a DS 2.0 manuscript entry. Options included: Full description of the manuscript, brief identifying description with link to institutional record, or open-ended other response.

3 other responses: "What level of metadata would be sufficient for a DS 2.0 manuscript entry for your collections?"

- "We are interested in contributing as much of our metadata as possible without raising the bar too high for small institutions with in-scope collections to participate."
- "A rich range of fields within the description that are searchable seems ideal to me, since we can link to a fuller description but the *searchability* of various metadata is of enormous benefit"
- "Brief is preferable as it decreases the need to update multiple repositories when new information is learned"

Figure 31: The three other responses from Figure 30, unabbreviated.

What level of digitization would be sufficient for a DS 2.0 manuscript entry for your collections?

Single images
8.6%

Cover-to-cover
28.6%

None
22.9%

Sample
40.0%
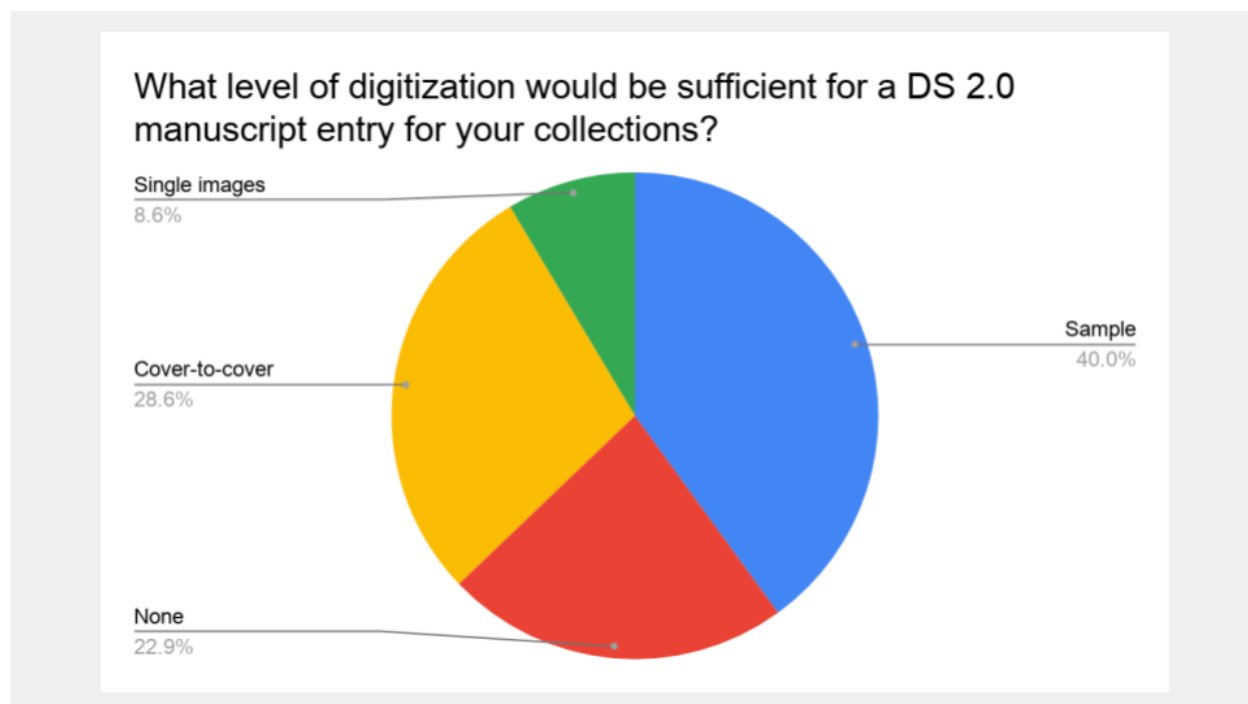
Figure 32: Level of digitization that would be sufficient for a DS 2.0 manuscript entry
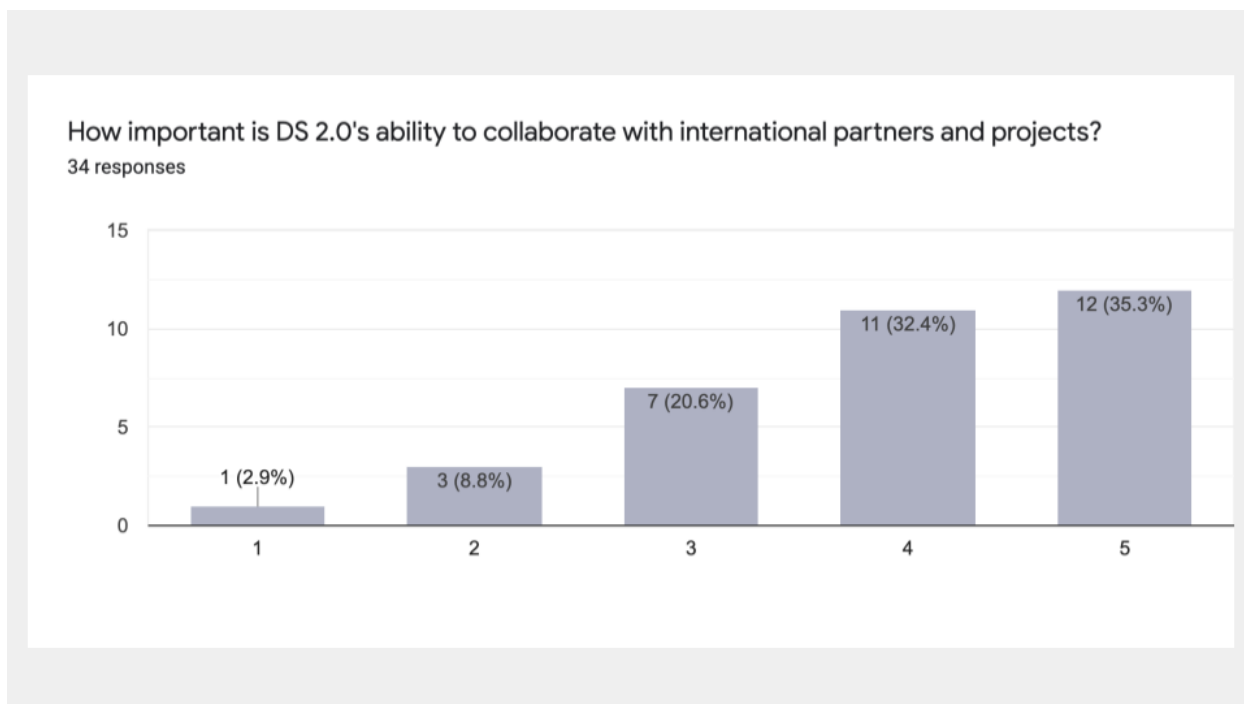
Figure 33: Ranking of importance of DS 2.0's ability to collaborate with international partners and projects. Scale ranged from 1 (not important) to 5 (extremely important)
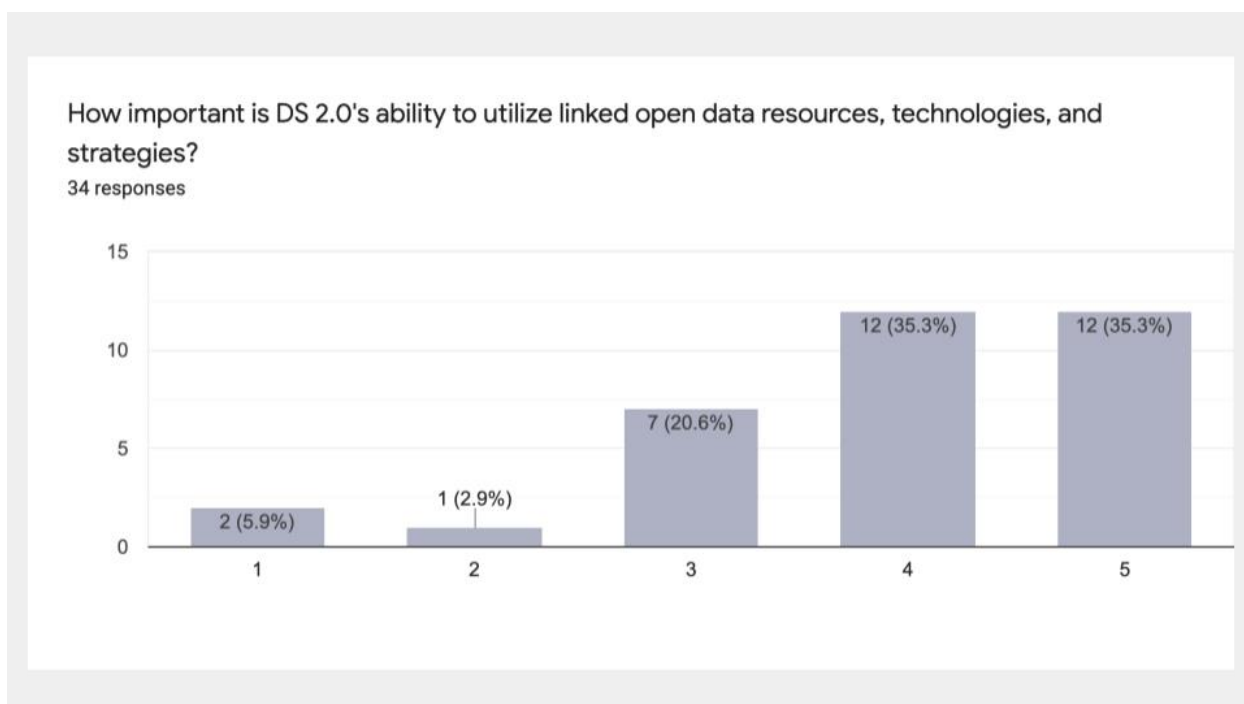
Figure 34: Ranking of importance of DS 2.0's ability to utilized linked open data resources, technologies, and strategies. Scale ranged from 1 (not important) to 5 (extremely important)

Summary of responses: "What other features and functions would you like to see in DS 2.0?"

- Authority control (6 responses)
- No cost to participate (3 responses)
- Easier ways to contribute (3 responses)
- Be more like e-codices (2 responses)
- Link to/integrate SDBM data (2 responses)
- LOD (2 responses)
- Better documentation for new users/contributors (2 responses)

Figure 35: Summary of responses to fill-in-the-blank question "What other features and functions would you like to see in DS 2.0?"

Appendix B. Planning Meeting Agenda

Agenda for DS 2.0 Virtual Planning Meetings

1pm-3pm ET, October 6, 8, 13, and 15

Core questions of the meetings:

What is a national union catalog of manuscripts in the 21st century? What are the possibilities? What is its scope? How is it sustained?

Breakout Group assignment: Describe your vision of DS 2.0

What is its mission? What is its scope? What data fields are mandatory? How will institutions or individuals contribute to DS 2.0? How will DS 2.0 serve both large and small institutions? What should DS 2.0 do for you as a researcher, instructor, or collection manager?

Daily Schedules:

Each day consists of two sections: presentations and breakout group work. Participation in a breakout group is not mandatory, but highly encouraged. All registered participants will be pre-assigned to a breakout group.

October 6: Digital Scriptorium and DS 2.0 Project Overview

1:00-1:45: Debra Cashion (Welcome and Intro to DS)

Lynn Ransom (DS 2.0)

1:45-2:10: Emma Thomson (DS Institutional Survey results)

2:10-2:15: Break

2:15-2:45: Breakout Groups

2:45-3:00: General Discussion/Wrap-up

October 8: Related Projects I

1:00-1:05: Welcome and Introductions

1:05-2:25: Sarah Noonan & Elizabeth Hebbard (Peripheral MSS Project)

Kelly Tuttle (Manuscripts of the Muslim Worlds)

Daniel Gullo (vHMML)

Christoph Flüeler (e-codices)

2:25-2:55: Breakout Groups

2:55-3:00: Wrap-up

October 13: Related Projects II

1:00-1:05: Welcome and Introductions

1:05-2:15: Benjamin Albritton (IIIF)

　　　　　Emma Thomson (Mapping Manuscript Migrations)

　　　　　Debra Cashion (Metascripta)

2:15-2:55: Breakout Groups

2:55-3:00: Wrap-up

October 15: Breakout Group Presentations: Your Vision for DS 2.0

1:00-1:05: Welcome and Introductions

1:05-1:35: Final meeting with Breakout Groups

1:35-1:45: Breakout Group 1 Presentation

1:45-1:55: Breakout Group 2 Presentation

1:55-2:05: Breakout Group 3 Presentation

2:05-2:15: Breakout Group 4 Presentation

2:15-3:00: General Discussion and Wrap-up

Appendix C. Planning Meeting Breakout Group Activity

Breakout Group Activity: Your Vision of DS 2.0

Breakout groups will meet during the first three DS 2.0 Planning Meeting sessions to develop their own vision of what DS 2.0 should be. Every registrant has been pre-assigned to a breakout group. You will be a member of this group for the entirety of the planning meetings. The groups will present their vision on the final day of the meeting (October 15).

The questions below are meant to foster discussion around the central issues concerning the development of a national union catalog. Use these questions to plan your breakout group's vision of DS 2.0. There are many questions listed in this document, and we do not anticipate that the groups will be able to answer all of the questions by the final day of the meeting. Your group may decide to focus on answering a specific set of questions that you consider most important. Each group will have 15 minutes to present. Your group's presentation can be as simple or complex as you like, as long as you explain what questions you discussed and how you arrived at your conclusions. We do not expect you to work on your presentation outside of the scheduled meeting times.

Your breakout group should consider the following questions:
- Which manuscripts should be included in DS 2.0?
  - What is the geographic or cultural scope? (All manuscript cultures around the world? Western European medieval and renaissance manuscripts? Something else?)
  - What about chronology? Will there be a strict date cutoff? Will the date cutoff be different for different materials?
  - Will DS 2.0 include only codices, or also non-traditional formats (rolls, tablets, etc.)?
- What should DS 2.0 records include?
  - Full catalog records, brief summaries, or something in-between? Why?
  - What data fields are mandatory?
- How will DS 2.0 balance the needs of large and small institutions?
  - What can DS offer large institutions that already host their own fully digitized collections?
  - What about smaller institutions that have uncataloged/undigitized mss?
  - What institutions should be included in DS 2.0?

- How will data get into DS 2.0? How will people contribute to DS 2.0?
    - Crosswalk pre-existing data? What if no pre-existing data exists?
    - Manual data entry? Who will do this work?
- How will DS 2.0 fit into the larger environment of digital mss tools and projects? For example:
    - Linked data
    - Authorities
    - Digitization
    - General interoperability
    - ???
- What should DS do for you as researcher, instructor, student, or collection manager? For example:
    - Provide/store images?
    - Provide/store metadata?
    - Facilitate sharing?
    - Open access?
    - Linkability?
    - ???
- Where should DS 2.0 live?
- How do we sustain DS 2.0?