IJDC | *General Article*

Community-based Curate-a-Thons to Enhance Preservation of Global Genetic Biodiversity Data: A Practical Case Study

Andrea L. Pritt
Pennsylvania State University Libraries, Harrisburg

Briana E. Wham
Pennsylvania State University Libraries, University Park

Rachel H. Toczydlowski
Northern Research Station, United States Forest Service

Eric D. Crandall
Pennsylvania State University, University Park

Abstract

Science, Technology, Engineering, and Mathematics (STEM) and Research Data Librarians collaborated with an international research team of conservation geneticists to create an instructional and practical guide combining genetic biodiversity initiatives and data curation. Over the course of two months, the academic librarians held multiple community-based Curate-A-Thons where an international group of students, researchers, librarians, and faculty researchers participated in tracking down publications and metadata for genomic sequence data, thus crowd-sourcing this effort of metadata enhancement. This article details the successful Curate-a-Thon design and implementation process; the openly available instructional materials created and used to host the Curate-a-Thons; and the challenges and successes of these community-based events.

Submitted 1 June 2023 ~ *Accepted* 6 October 2023

Correspondence should be addressed to Andrea Pritt, Madlyn L. Hanes Library, Penn State Harrisburg, 351 Olmsted Drive, Middletown, PA 17057. Email: alp5088@psu.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The sheer volume at which data are being created globally is both astonishing and exponentially growing. Additionally, there are increasing expectations among funders (Boehm et al., 2023), publishers (Jones, Grant, & Hrynaszkiewicz, 2019), and academic institutions (Llebot & Castillo, 2023) that data be managed and openly shared in support of research reproducibility and transparency as well as to enable reuse (Briney, Goben, & Zilinski, 2017). Yet frequently data that are shared (e.g., in online repositories and published scientific papers) are lacking the contextual metadata necessary for interpretation and reuse (Sabot, 2022; Pope et al., 2015). Thus, proper data curation as well as data curation outreach and education are necessary for both quantitative and qualitative data across all disciplines, including social sciences, the humanities, and science, technology, engineering, and mathematics (STEM), to guarantee that the data are findable, accessible, interoperable, and reusable (FAIR) and preserved for the long term (Wilkinson et al. 2016).

In particular, data that describe the genetic diversity of species are currently not very FAIR (Toczydlowski et al., 2021). Genetic diversity is the most fundamental form of biodiversity: in the same way that the number of species in an ecosystem can support ecosystem health, the number of possible genotypes within a species can support its evolutionary health and ability to adapt to a changing environment (Raffard et al., 2019). The sharing of genomic data has been broadly established within the genetics community, with broad cultural acceptance and support from journals and funders for widespread use of established, data-type repositories (Crandall et al., 2023; Toczydlowski et al., 2021; Byrd et al., 2020; Pope et al., 2015). The most prominent of these repositories are the Sequence Read Archive and GenBank databases within the International Nucleotide Sequence Database Collaboration (INSDC).[1] Note the INSDC is a partnership of the National Canter for Biotechnology Information (NCBI), DNA Data Bank of Japan (DDBJ), and European Nucleotide Archive at the EMBL European Bioinformatics Institute (EMBL-EBI) (Cochrane et al., 2016).

However, recent research has shown that much of these openly available genomic data are lacking the spatiotemporal metadata necessary for reuse in ecological, evolutionary, and conservation studies, as well as in the estimation of biodiversity metrics more specifically (Crandall et al., 2023; Toczydlowski et al., 2021; Riginos et al., 2020; Pope et al., 2015). In other words, we do not know when or where in the world many of the genomic sequences in INSDC were collected from. Genomic records for wild species relevant for biodiversity studies within the INSDC only included both geospatial coordinates and collection years in 13% of records, geospatial coordinates in 17% of records, and collection years in 40% of records (Toczydlowski et al., 2021). Thus, this gap in associated metadata emphasizes the need for accurate and timely metadata curation of these genomic data to enable conservation practitioners and geneticists working on biodiversity initiatives to reuse these data (Toczydlowski et al., 2021; Crandall et al., 2023).

Multiple efforts have recently been started to support FAIR archiving of genomic metadata. An international research team developed the Genomic Observatories Metadatabase (GEOME) to centralize and standardize genomic metadata archiving. GEOME eases the development, capture, and linkage of metadata for biological samples and their associated genetic sequences stored in the INSDC (Deck et al., 2017; Riginos et al., 2020). In a related effort, researchers designed and ran a distributed datathon to determine how much of the missing metadata in the INSDC could be recovered from external sources (e.g., associated scientific papers and authors – (Crandall et al., 2023). The 12 paid, part-time graduate students that comprised the first genomic metadata curation datathon were able to successfully recover and restore spatiotemporal metadata for 78% of the 561 datasets that were addressed (Crandall et al., 2023). Readers interested in a detailed description of the background, scope, and workflow of the first genomic metadata curation datathon, and the scientific implications of it, are referred to

[1] https://www.insdc.org/

Toczydlowski et al. (2021, state of metadata in INSDC) and Crandall et al. (2023, metadata recovery efforts).

Academic librarians are uniquely positioned to further advance metadata preservation efforts. First, academic librarians have a demonstrated track record of collaborating with faculty research teams in productive ways. These collaborations have included writing systematic reviews (Lee et al., 2022; Spencer and Eldredge, 2018; Foutch, 2016), improving existing course assignments to enhance students' information literacy skills (Becker et al, 2022; Douglas & Rabinowitz, 2016), and securing grant funding (Lehner-Quam, 2022). In addition to these efforts, librarians have the essential skills required to lead, and assist with, data management including data curation, metadata management, and navigating the research data lifecycle (Lee & Stvilia, 2017; Pouchard, 2015). Finally, academic librarians frequently develop and lead outreach programming to provide scholarly opportunities outside of the classroom. Crowdsourcing events such as transcribe-a-thons and Wikipedia edit-a-thons, outreach tools that takes advantage of the efficiency of the multitude (Ellis, 2014), provide an excellent model for contributing to metadata preservation. Crowdsourcing outreach programming also has the added benefits of enhancing patron engagement, sense of community, diversity of viewpoints, and information literacy (Kasten-Mutkus, 2020). They also provide academic librarians another avenue through which to make meaningful contributions at their institutions.

Much has been published about academic institutions and libraries conducting edit-thons and transcribe-a-thons spanning several unique disciplines (PSU News, 2023; Douglass Day, 2023; Littlejohn et al., 2021; Bridges and Dowell, 2020; Di Lauro, 2020; Mareca & Bordel, 2019; Weiner et al, 2019; Sliger Krause et al 2017). The purpose of these edit-a-thons and transcribe-a-thons is often similar: to introduce editors (typically students) to collaborative environments with the goal of creating new knowledge and/or improving the quality and accessibility of existing knowledge. Often led by Higher Education Institutions (HEIs) in the United States and internationally, improving the quality of publicly available consumer health information, improving the gender balance of existing Wikipedia articles, and enhancing scientific articles are a few examples of this type of work (Di Lauro, 2020; Mareca & Bordel, 2019; Weiner et al, 2019). Edit-a-thons are increasingly being offered outside of the formal classroom as academic librarians are becoming more involved. For example, librarians at Pennsylvania State University host annual transcribe-a-thons on Douglass Day (February 14),[2] as a collective action during Black History Month and Wikipedia edit-a-thons focused on Native American women activists and environmentalists (Douglas Day, 2023; Molnar, 2023). Edit-a-thons such as these provide participants with an opportunity to work collaboratively in a digital learning environment (Littlejohn et al., 2021; Bridges & Dowell, 2020; Sliger Krause et al, 2017).

This article details how librarians at Pennsylvania State University, an R1 public university in Pennsylvania (Carnegie Classification of Institutions of Higher Education [3]), collaborated with STEM researchers to adapt the structure of other crowd-sourced events to increase participation and offer data curation training and build awareness about the value of good metadata for data discoverability, interoperability, and reusability. More specifically, this article outlines a recent case study of how crowd-sourced, community-based Curate-a-Thon events were developed and hosted as part of the GEODE: A Genomic Observatories Diversity Explorer project to create a unique, innovative, and engaging collaborative learning experience for a diverse audience including undergraduate and graduate students, researchers, and librarians with varying experience levels with data curation and understanding of genetic sequence data and biodiversity. The practical design and development process of running multiple, crowdsourced Curate-a-Thons and the outcomes are outlined with additional focus on the challenges and successes of the events.

[2] https://douglassday.org/
[3] https://carnegieclassifications.acenet.edu/institution/the-pennsylvania-state-university/

Approach

Designing the Curate-a-Thon

The overarching goal of developing and hosting the Curate-a-Thons was to support genetic biodiversity initiatives and the ability to incorporate biodiversity data into large scale conservation policy decisions by improving the Findability, Accessibility, Interoperability, and Reusability (FAIR principles) of genetic sequence data (Wilkinson et al., 2016). More specifically, the Curate-a-Thons aimed to: (1) encourage community engagement with genetic biodiversity data and metadata, (2) increase the accessibility and preservation of genetic data for reuse by more communities, (3) share knowledge about genetic biodiversity and data curation with interested participants, and (4) improve the efficiency of the datathon, in its second iteration, by adopting a crowdsourced approach.

To these aims, during the summer of 2022, we scraped the Sequence Read Archive (SRA), the INSDC's repository dedicated to second-generation sequence data, to generate an initial list of BioProjects (datasets) that would be curated during the Curate-a-Thons. Curate-a-Thon participants were assigned BioProjects (see description below) and searched for published papers associated with their assigned BioProjects. A BioProject is a dataset of genetic sequences associated with a project which often includes genetic sequence data files and sample information. By reading these papers, participants determined whether the BioProject's genetic data were relevant for biodiversity initiatives (i.e., were they sampled from wild populations), and identified whether spatial or temporal metadata were present in the published paper or associated documents (supplemental materials, data repository deposit, etc.) (Figure 1).

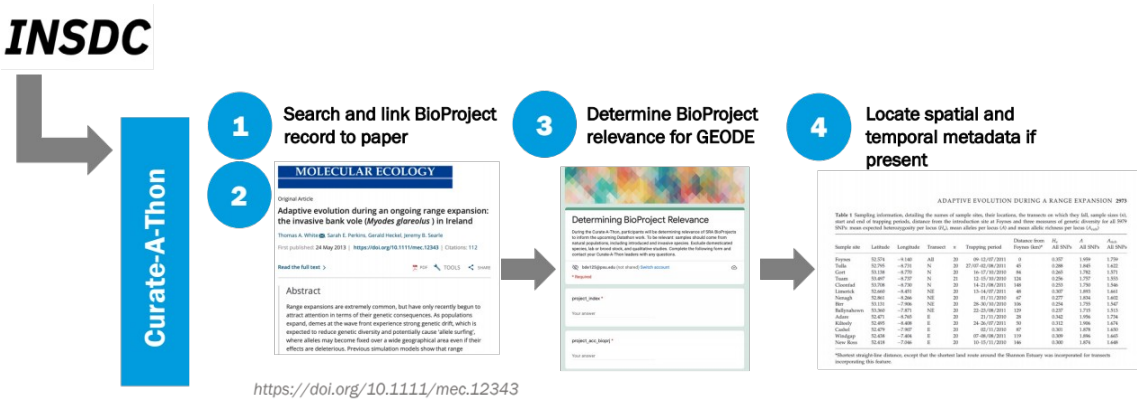


Figure 1. Curate-a-Thon workflow to crowd-source the enhancement or recovery of metadata relevant to genetic biodiversity efforts missing from genetic sequence data stored in the INSDC.

Based on the experiences of researchers involved in the first genomic metadata curation datathon, it was apparent that there was a need to increase the efficiency of the workflow. Specifically, records accessioned to the INSDC SRA do not indicate whether the sequenced tissue sample came from a wild population or from a domesticated or otherwise human-managed population. The distinction is important because only genomic data from the former category are relevant to efforts to conserve wild genetic diversity. Curators of the first genomic metadata curation datathon spent a lot of time making this determination, which reduced the amount of time available for actual accessioning of the metadata associated with relevant (wild) datasets.

We were inspired by the edit-a-thon/transcribe-a-thon outreach programming efforts and believed that this format could be adapted to support and improve the efficiency of genomic metadata curation efforts. Thus, in this second iteration, we added crowd-sourced events, Curate-a-Thons, prior to the genomic metadata curation datathon to initially associate articles,

determine relevance, and discover spatial/temporal metadata. In adapting this outreach programming format, we believed we would be able to offer participants a beneficial hands-on learning experience in information literacy and data curation, as well as exposure to the issues of missing contextual metadata for biodiversity initiatives, while also progressing the work of the research project through crowd-sourcing efforts.

Additionally, we recognized that this would be an opportunity to engage a broader and more diverse group of participants about the importance of rich metadata for reuse. Therefore, in designing the activities of the Curate-a-Thons, we selected activities which were well-suited for a broad and diverse group of participants including undergraduate and graduate students, researchers, librarians of varying expertise and experience with data curation, genetic sequence data, and biodiversity research. Though this research is situated deeply within genetic biodiversity initiatives and digital curation, Curate-a-Thon participants were not required to have any previous experience in the sciences or curatorial processes. In fact, new and relatively inexperienced curators were highly encouraged to participate, as there is ample cross-checking and quality control built into the overall process for metadata remediation (Crandall et al., 2023). The team felt this was a good opportunity to collaborate with a diverse group of participants that may otherwise not have the opportunity to engage with global genetic biodiversity initiatives.

In the following sections, we describe in more detail how we developed and hosted the Curate-a-Thons, including the development of materials and logistics, and reflect on the challenges and successes of these events.

Developing the Curate-a-Thon Materials

We developed a set of instructional materials for the Curate-a-Thon. To this end, we created groups of materials designed for different purposes and users. When completed, materials were created for:

- Advertising and marketing the Curate-a-Thon which includes:
    - Suggested email templates for a Call for Participants/Curators,
    - Suggested email template for a Call for Curate-a-Thon Hosts, and
    - Digital promotional flyers.
- Curate-a-Thon Hosts which includes:
    - Written instructions on how to successfully run your own Curate-a-Thon,
    - Necessary day-of Curate-a-Thon materials including informational presentation slides, links to pre-filled BioProject Google Sheets for Curate-a-Thon participants, and
    - Follow-up materials regarding incentives for active participants.
- Curate-a-Thon Participants which includes:
    - Curate-a-Thon Participant Guide hosted via GitHub Pages,
    - Designated Google Sheet for each individual participant, and
    - Set of pre-filled BioProject Google Forms that were specific to each participant.

The Curate-a-Thon Participant Guide included written information; step-by-step written instructions accompanied by screenshots and brief video tutorials and has been published openly (Pritt et al, 2022). The same information was provided in different ways to best accommodate

the differing learning styles and expertise levels of participants. The guide also provided introductory information on getting started, which included a data curation primer and terms to know. We anticipated that reading through the participant guide, from start to finish, would take approximately 30 minutes. This includes the time spent watching the brief video tutorials if the curator chose to. Additionally, we developed a specific, detailed metadata curation protocol for all participants to follow. A Frequently Asked Questions (FAQ) section was added to anticipate participant questions. The participant guide served as a one-stop-shop for participant needs both before and during the Curate-a-Thon. We detail when and how we provided this guide to participants below (see *Curate-a-Thon Logistics*).

Instead of using a master spreadsheet of all BioProjects for all curators to work on simultaneously, we created a pre-filled Google Form for each individual BioProject. The pre-filled Google Form contained metadata extracted from the SRA regarding the BioProject and empty fields for curator responses regarding links to associated papers, relevance, and presence/absence of spatial and/or temporal information, as well as their reasoning for each determination. Asking participants to provide rationale for their responses served as an initial quality control assurance step and utilizing pre-filled Google Forms instead of a master sheet for data collection was implemented to minimize errors. The pre-filled Google Forms also included SRA and BioProject-specific metadata fields that participants were instructed not to edit.[4] Each participant was assigned one Google Sheet for the Curate-a-Thon which included five BioProjects in need of curation and links to five separate pre-filled BioProject Google Forms.

Curate-a-Thon Logistics

As an international team of conservation biologists, research scientists, academic librarians, and information technology experts – all with varying backgrounds, degrees, and expertise levels – it was important to our team to seek out Curate-a-Thon participants, internationally, from a wide variety of academic backgrounds and experiences, as well. To that end, we decided to host the Curate-a-Thons virtually to best accommodate participants from all over the world, in any time zone. A second quality assurance step was that at least two data curators would curate each BioProject. If each curator determined the relevance to be different (one metadata curator determined the BioProject to be "relevant" and the other curator determined it to be "not relevant"), a third curator would later review the BioProject to split the tie and make a final decision.

In total, there were seven Curate-a-Thons hosted by the authors: three in November 2022 and four in January 2023. Each Curate-a-Thon was three to four hours in length and participants were welcome to stay for the entire event or drop in/out at times that were convenient for them. Five events were held in the late afternoon Eastern Standard Time to accommodate different time zones. The remaining two events were held in the morning Eastern Standard Time. Participants were asked to join the event on the half-hour or hour marks, to make onboarding multiple participants at the same time easier (and less disruptive to other participants). Registration was required for each event and all Curate-a-Thons were held virtually via Zoom.

Curate-a-Thons were advertised at least one month before the scheduled event date. Advertisements were sent to a variety of email listservs targeting genetic biodiversity organizations; college and university students; science academic librarians; and research data librarians. In addition to targeted email listservs, undergraduate biology students from Pennsylvania State University, with a grade point average of 3.0 or higher, were specifically invited to participate. To encourage participation, $10.00 Amazon electronic gift cards were offered to active participants who stayed for at least one hour and/or curated at least five BioProjects. If a participant stayed longer and/or curated additional BioProjects, they were eligible to receive additional gift cards while supplies lasted.

[4] Readers interested in reviewing metadata curation protocol and other materials are encouraged to review the Curate-a-Thon Participant Guide, openly available, for examples: https://bdezray.github.io/Geode-Curate-A-Thon/

The day before a Curate-a-Thon was scheduled, we sent all registrants an introductory welcome email. This email included the Curate-a-Thon Zoom link, instructions on when to join, and a suggestion to read through the Curate-a-Thon Participant Guide before attending to gain a better understanding of the project and to make the onboarding process run smoothly. At the start of each event, and every subsequent half-hour and hour, new participants were onboarded. The onboarding process included a short, three-minute, verbal explanation of how the event would run and information about how the assigned Google Sheets would be shared with participants. The brief explanation included six steps, to be completed in order:

1. Read GEODE Curate-a-Thon Participant Guide
2. Open your assigned Google sheet to access your BioProjects
3. Follow protocol directions to search for associated paper and locate metadata
4. Fill in and submit BioProjects Google Form
5. Updated Google Sheet to reflect status (submitted yes/no) of BioProject
6. Begin next BioProject listed in your Google Sheet

Participants were encouraged to ask questions via the Zoom chat and/or audibly with their microphone.

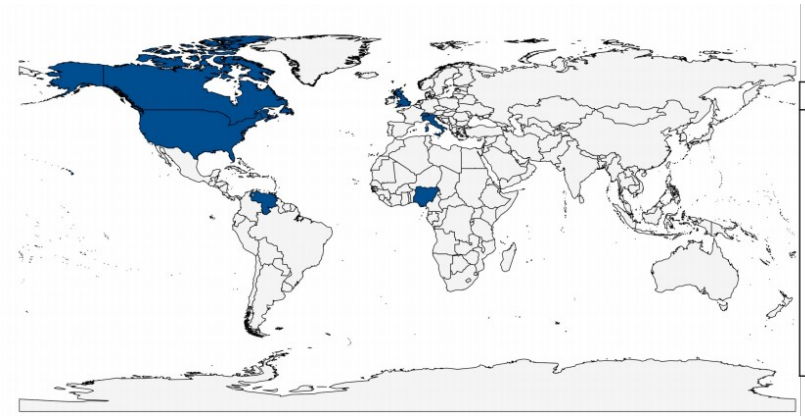
Results and Discussion

Curate-a-Thon by the Numbers

There was a total of 125 participants for a total of seven Curate-a-Thons, which led to the curation of 251 BioProjects. Each BioProject was curated by at least two data curators. Table 1 provides a detailed breakdown of participants. Graduate student / researcher participant type includes participants who indicated their affiliation with an academic institution. Community member participant type includes participants who did not indicate their affiliation with an academic institution. Figure 2 provides a detailed breakdown of participants by geographic location.

Table 1. Breakdown of Curate-a-Thon registrants by participant type.

Participant Type	Number of Participants
Graduate student / researcher	71
Undergraduate student	21
Librarian	19
Community member	12
United States Governmental Dept.	2



Geographic Location	# of Participants
United States	112
Canada	3
Italy	2
Nigeria	2
Switzerland	2
Iran	1
Scotland	1
United Kingdom	1
Venezuela	1

Figure 2. Geographic distribution of Curate-a-Thon participants denoted in blue on the map on the left with participant counts by geographic location in a table on the right.

Challenges

Prior to running the Curate-a-Thons, we had no direct experience developing, organizing, or hosting a community-based, crowdsourced event such as this. Being able to reference the successful first genomic metadata curation datathon was invaluable (Crandall et al., 2023), but we faced several challenges during the material creation process as well as during the events themselves.

Before hosting the Curate-a-Thons, the process of creating a pre-filled Google Form was new to us and though it worked out well in the long run, there was a brief learning curve in the process. As the pre-filled Google Form was being developed, the research team was making changes to what information should be added to the form, what information should be pre-filled, and what information should the Curate-a-Thon participants be responsible for locating. Due to these ongoing discussions, building the pre-filled Google Form link took several days and iterative attempts to perfect.

To create the Curate-a-Thon Participant Guide, we were able to lean on the materials previously created for the first genomic metadata curation datathon, but with the goal of improving upon and enhancing them. Because the Curate-a-Thon was designed to be phase one of the process in the second iteration, the challenge was to develop and write entirely new sections to address aspects specific to the Curate-a-Thon process. The pre-filled BioProject Google Forms were not used at all in the first genomic metadata curation datathon conducted in 2020 and required ample documentation to explain. There was the challenge of writing documentation that would be easily understood by novices as well as experienced participants. For accessibility purposes, short video tutorials explaining several aspects of the Curate-a-Thon were recorded (and rerecorded) and embedded into the Participant Guide. The guide took several weeks to write, tweak, format, and redesign and each version of the written documentation was shared with team members for edits and suggestions. While the guide was being written, we worked through the metadata curation process and captured screenshots to further enhance the documentation. As with any highly collaborative research project, there were many simultaneously moving parts which took a coordinated effort to manage and work through.

Another challenge was the necessary time and organizational effort that it took to create BioProject-specific Google Sheets, add the unique pre-filled BioProject Google Form links to each sheet, and assign each sheet to a specific participant. This was done to randomize the participants with corresponding BioProjects to ensure BioProjects were being curated by at least two participants. After each Curate-a-Thon, incomplete pre-filled forms needed to be cleared

and then reassigned to future participants which also took a considerable amount of time and effort.

The biggest challenge while hosting was fielding a large number of simultaneous questions. We did our best to answer participant questions in the order they were received, but we were often messaging each other privately on the side to clarify and/or confirm our answers to limit confusion and errors. Some questions made it clear that participants had not reviewed the Participant Guide documentation and/or the FAQ section of the guide. Another anticipated challenge was the varying levels of knowledge and experience of participants and how prior experience informed the questions that were asked Simultaneously answering questions during live, synchronous working sessions – from two different locations – was confusing at times. The Curate-a-Thons would have benefitted from additional team members in the sessions for quicker responses, but ultimately all participant questions were answered.

Minor technology issues occurred during the live events but were limited to not having edit access to assigned Google sheets and general Zoom microphone and camera connection issues. Additionally, all the live events were held with priority to Eastern Standard Time working hours and entirely run in English. The time zone differences made it somewhat difficult for participants to join live (which led to a pivot in offering this as asynchronous work – more about that later) and though most, if not all, participants knew English, it may not have been their first language. These challenges led to minor back-and-forth discussions, where participants were offered to join a private Breakout Room in Zoom for more personalized assistance.

Successes

Though challenges arose at different points during the process, all four Curate-a-Thon goals were successfully met. As seen in Table 1, the events brought together a wide variety of participants; some with scientific data experience and some without it. Many of the participants self-identified as having experience with data curation, bioinformatics research, as well as professional experience of working in a scientific laboratory. Others identified as academic librarians or community members who worked in STEM-adjacent or even non-STEM fields. Data curators' experience level ranged from community members (those without an institutional/university affiliation) to graduate researchers to two participants from the United States Department of Agriculture. No formal assessment of the Curate-a-Thons was conducted but the feedback was overwhelmingly positive. Participants shared the following feedback with Curate-a-Thon hosts after they curated BioProjects:

> "Thank you for this amazing opportunity! I wouldn't mind to complete some more in my down time, as a way to kind of let my mind focus on something entirely new and help refresh my teaching brain."
>
> "This is super fun, glad to see you all doing this!"
>
> "Thank you. This was an interesting experience especially since I am often given the task to upload sequence info to the SRA at NCBI."
>
> "Thank you so much for hosting and giving us opportunities to participate. It was a new experience to me and I learnt a lot!"
>
> "I was only able to submit two forms, but really enjoyed the practice. I curate some datasets for my institution's data catalog, so it was great to get the opportunity to familiarize myself with some of the genomic/genetic terms. Thanks again for hosting a very cool (and well organized!) curate-a-thon."
>
> "Thanks so much for organizing this event! I got to explore research on Mongolian horses, yeast, Australian birds and Japanese abalone! It was fun to combine my

> librarian skills of hunting down articles and my science skills of (somewhat) interpreting a Methods section and Supp Info."
>
> "Kudos on the protocol and the page btw, it's very well structured and it's obvious that a lot of work went into it!"

Unsolicited feedback such as these comments were indicators that goal one was met, and the Curate-a-Thons were well received by all. The previously noted challenge of the volume of questions the instructors received during the live, synchronous event led to an unforeseen benefit. These new questions were incorporated into the Participant Guide's FAQ to build a more comprehensive and inclusive document. By locating missing spatiotemporal metadata, the team will be able to increase both the accessibility and preservation of this data during phase two, the datathon. By determining BioProject relevance, and ultimately locating the missing metadata, Curate-a-Thon participants were able to assist in successfully meeting goal two.

As STEM and Research Data Management librarians we were able to positively share our knowledge and expertise, as STEM data curators, with all interested participants meeting goal three. Perhaps most importantly, the Curate-a-Thon events included diverse attendance and though we would have liked to see more international participation, having participants from eight international countries was a success. Born out of the challenge of differing time zones was an opportunity for participants to work asynchronously at a time that was more suitable for them. This was another unforeseen success which led to a handful of participants participating offline when they were able; reaching out to the Curate-a-Thon instructors via email to ask questions and/or to confirm the completion of metadata curation. Those who participated asynchronously often asked to curate additional BioProjects, which we happily obliged.

Incentives in the form of $10.00 USD Amazon electronic gift cards were given to participants who curated at least five BioProjects and/or participated for at least one hour. If a participant curated additional BioProjects (or stayed for multiple hours) they were given additional incentives. In addition to the incentives, undergraduate students from the Pennsylvania State University were also offered extra credit for a 400-level Biology course. Originally designed to be the main driver of participation in a Curate-a-Thon, multiple participants began declining the incentives stating they were happy to continue to participate and were no longer interested in receiving supplemental gift cards. This allowed new participants to be recruited, and all participants received at least one gift card. In total, there were 251 BioProjects curated by at least two participants. Though this was only approximately 10% of the total BioProjects that needed to be curated, we were pleased with the outcome. Phase two of this project, the datathon, moves the work from volunteer, incentive-backed participation into paid, part-time data curator positions. Curate-a-Thon participants were highly encouraged to apply for these paid, part-time positions and one final success was that three of the ten data curators who were hired, had direct Curate-a-Thon metadata curation experience.

Conclusion

Overall, we felt that this was a team success and would continue hosting additional Curate-a-Thons in the future. From start to finish this process gave us the opportunity to build upon our pre-existing STEM and data librarianship skills while collaborating with a new and diverse community. The Curate-a-Thon events offered an opportunity to locate missing spatiotemporal metadata and the second iteration of the datathon will continue that effort to enhance the preservation of this important global genetic biodiversity data. We hope that by sharing our experiences; our challenges, and our successes, that others will build upon this work and engage new communities in the important, and necessary, metadata curation work.

The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or United States Government determination or policy.

Acknowledgements

In 2022, this project received a grant from the GEO-Microsoft Planetary Computer program, entitled *GEODE: A Genomics Observatories Diversity Explorer* with Crandall as Principal Investigator. Funds provided by this grant were used to purchase the incentives for Curate-a-Thon participants.

This research was supported in part by the U.S. Department of Agriculture, Forest Service.

References

Becker, J. K., Bishop Simmons, S., Fox, N., Back, A., & Reyes, B. M. (2022). Incentivizing information literacy integration: A case study on faculty–librarian collaboration. *Communications in Information, 16*(2), 167–181. doi:10.15760/comminfolit.2022.16.2.5

Boehm, R. I., Condon, P. B., Calkins, H., Petters, J., & Woodbrook, R. (2023). Analysis of US Federal Funding Agency Data Sharing Policies. *International Journal of Digital Curation, 17*(1), 18. doi:10.2218/ijdc.v17i1.791

Bridges, L. M. & Dowell, M. L., (2020). A perspective on Wikipedia: Approaches for educational use. *The Journal of Academic Librarianship, 46*(1), 102090. doi:10.1016/j.acalib.2019.102090

Briney, K., Goben, A., & Zilinski, L. (2017). Institutional, Funder, and Journal Data Policies. In Lisa Johnston (Ed.), *Curating Research Data: Volume One*. ACRL.

Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X., & Greene, C. S. (2020). Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics*, *21*(10), 615-629. doi:10.1038/s41576-020-0257-5

Cochrane, G., Karsch-Mizrachi, I., Takagi, T., & Sequence Database Collaboration, I. N. (2016). The international nucleotide sequence database collaboration. *Nucleic acids research*, *44*(D1), D48-D50. doi:10.1093/nar/gkv1323

Crandall, E. D., Toczydlowski, R. H., Liggins, L., Holmes, A. E., Ghoojaei, M., Gaither, M. R., Wham, B. E., Pritt, A. L., Noble, C., Anderson, T. J., Barton, R. L., Berg, J. T., Beskid, S. G., Delgado, A., Farrell, E., Himmelsbach, N., Queeno, S. R., Trinh, T., Weyand, C., … Toonen, R. J. (2023). Importance of timely metadata curation to the global surveillance of genetic diversity. *Conservation Biology*, e14061. doi:10.1111/cobi.14061

Deck J., Gaither M. R., Ewing R., Bird C.E., Davies N., Meyer C., et al. (2017) The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biology, 15*(8): e2002925. doi:10.1371/journal.pbio.2002925

Di Lauro, F. (2020). 'If it is not in Wikipedia, blame yourself:' edit-a-thons as vehicles for computer supported collaborative learning in higher education. *Studies in Higher Education, 45*(5), 1003-1014. doi:10.1080/03075079.2020.1750191

Douglas, V. & Rabinowitz, C. (2016). Examining the relationship between faculty-librarian collaboration and first-year students' information literacy abilities. *College & Research Libraries, 77*(2), 144-163. doi:10.5860/crl.77.2.144

Ellis, S. (2014). A history of collaboration, a future in crowdsourcing: Positive impacts of cooperation on British librarianship. *Libri*, *64*(1), 1-10. doi:10.1515/libri-2014-0001

Foutch, L. J. (2016). A new partner in the process: The role of a librarian on a faculty research team. *Collaborative Librarianship, 8*(2), 80-83.

Jones, L., Grant, R. and Hrynaszkiewicz, I. (2019). Implementing publisher policies that inform, support and encourage authors to share data: two case studies. *Insights: the UKSG journal*, 32(1), 11. doi:10.1629/uksg.463

Kasten-Mutkus, K. (2020). Programming as Pedagogy in the Academic Library. *portal: Libraries and the Academy 20*(3), 425-434. doi:10.1353/pla.2020.0023

Lee, D. J., & Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PloS One*, *12*(3), e0173987. https://doi.org/10.1371/journal.pone.0173987

Lee, M. S., Hughes, A., Lockmiller, C., Day, A., Brown, M., & Jenson, R. (2022). Working together: How academic librarians can help researchers prepare for a grey literature search for systematic reviews involving minoritized populations. *The Journal of Academic Librarianship*. doi:10.1016/j.acalib.2022.102626

Lehner-Quam, A. (2022). Diversifying and transforming a public university's children's book collection: Librarian and teacher education faculty collaboration on grants, research, and collection development. *Collection Management, 47*(2-3), 157-178. doi:10.1080/01462679.2021.1958400

Littlejohn, A., Hood, N., Rehm, M., McGill, L., Rienties, B., & Highton, M. (2021). Learning to become an online editor: the editathons as a learning environment. *Interactive Learning Environments, 29*(8), 1258-1271. doi:10.1080/10494820.2019.1625557

Llebot, C. & Castillo, D. J., (2023) Are institutional research data policies in the US supporting the FAIR Principles? A content analysis. *Journal of eScience Librarianship*, *12*(1), 1-16. doi:10.7191/jeslib.614

Mareca, M. P. & Bordel, B. (2019). The educative model is changing: toward a student participative learning framework 3.0-editing Wikipedia in the higher education. *Universal Access in the Information Society, 18*, 689-701. doi:10.1007/s10209-019-00687-6

Molnar, B. (2023, March 23). *Libraries to host virtual Wikipedia editathon focusing on Native American women*. https://www.psu.edu/news/impact/story/libraries-host-virtual-wikipedia-editathon-focusing-native-american-women/

Pope, L. C., Liggins, L., Keyse, J., Carvalho, S. B., & Riginos, C. (2015). Not the time or the place: the missing spatio-temporal link in publicly available genetic data. *Molecular Ecology, 24*(15), 3802-3809. doi:10.1111/mec.13254

Pouchard, L. (2015). Revisiting the data lifecycle with big data curation. *International Journal of Digital Curation, 10*(2), 176-192. doi:10.2218/ijdc.v10i2.342

Pritt, A. L., Wham, B. E., Toczydlowski, R. H., & Crandall, E. D. (2022). GEODE Curate-a-Thon Participant Guide. https://bdezray.github.io/Geode-Curate-A-Thon/

Raffard, A., Santoul, F., Cucherousset, J., & Blanchet, S. (2019). The community and ecosystem consequences of intraspecific diversity: A meta-analysis. *Biological Reviews, 94*(2), 648–661. doi:10.1111/brv.12472

Riginos, C., Crandall, E. D., Liggins, L., Gaither, M. R., Ewing, R. B., Meyer, C., ... & Deck, J. (2020). Building a global genomics observatory: Using GEOME (the Genomic Observatories Metadatabase) to expedite and improve deposition and retrieval of genetic data and metadata for biodiversity research. *Molecular Ecology Resources*, *20*(6), 1458-1469. doi:10.1111/1755-0998.13269

Sabot, F. (2022). On the importance of metadata when sharing and opening data. *BMC Genomic Data*, *23*(1), 79. doi:10.1186/s12863-022-01095-1

Sliger Krause, R., Rosenzweig, J., & Victor, P. J. (2017) "Out of the vault: Developing a Wikipedia edit-a-thon to enhance public programming for university archives and special collections," *Journal of Western Archives*: Vol. 8: Iss. 1, Article 3. doi:10.26077/6730-6504

Spencer, A. J. & Eldredge, J. D. (2018). Roles for librarians in systematic reviews: a scoping review. *Journal of the Medical Library Association, 106*(1), 46-56. doi:10.5195/jmla.2018.82

Toczydlowski, R. H., Liggins, L., Gaither, M. R., Anderson, T. J., Barton, R. L., Berg, J. T., Beskid, S. G., Davis, B., Delgado, A., Farrell, E., Ghoojaei, M., Himmelsbach, N., Holmes, A. E., Queeno, S. R., Trinh, T., Weyand, C. A., Bradburd, G. S., Riginos, C., Toonen, R. J., & Crandall, E. D. (2021). Poor data stewardship will hinder global genetic diversity surveillance. *Proceedings of the National Academy of Sciences, 118*(34), e2107934118. doi:10.1073/pnas.2107934118

Weiner, S. S., Horbacewicz, J., Rasberry, L., & Bensinger-Brody, Y. (2019). Improving the quality of consumer health information on Wikipedia: Case series. *Journal of Medical Internet Research, 21*(3), e12450. doi:10.2196/12450

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. doi:10.1038/sdata.2016.18