Arts Access Datathon
Datathon Toolkit

Scott Reed, Datathon Archivist

This toolkit is intended to provide resources, recommendations, and communication templates for organizing a similar datathon, based on our experience with the Arts Access Datathon that took place on April 22, 2017. It covers the following topics:

1. Planning and managing the event
2. Compiling datasets
3. Designing the contest
4. Archiving and preserving event materials, media, and datasets

1. Planning and Managing the Event

The Arts Access Datathon leadership group began planning for the event approximately a year in advance. The core leadership team consisted of three (later expanded to four) members from the two co-hosting organizations. An informal memorandum of understanding (see a generic MOU template here) was drafted that outlined each organization's roles and responsibilities. We used a combination of Google Docs and Sheets to coordinate and document our planning process.

Steering Committee

The leadership team brought together a diverse group of stakeholders to serve as a steering committee, as well as a larger advisory committee. Diversity included race and ethnicity, gender, age, seniority in their field, LGBTQIA identity, and geography. In addition, we were careful to include people with different levels of knowledge and perspectives on the arts, access, and data. We intentionally included people with very limited formal experience working with data. We also made sure to include artists. This diversity of viewpoints and experiential backgrounds at times slowed our planning process as we built a shared understanding of what we sought to achieve, but the benefits made it worthwhile. It deepened our work and helped us to ensure an event that would be relevant and welcoming to a broad range of people.

Steering committee members met on a monthly basis, providing leadership and resources for the event. The advisory committee was loosely assembled via email and their attendance was not mandatory at any point. Steering committee members were primarily engaged in planning as well as securing in-kind resources and volunteers.

Registration Fee

After much discussion among the steering committee, the steering committee decided to charge a nominal fee to attend. Significant financial support for the Arts Access Datathon came from Sotheby's Art Institute at Claremont Graduate University, which also provided space for the event at no charge. Arts for LA provided support for registration and labor. While we secured other donations of funds, goods, and services, we did not seek the kind of major corporate underwriting that often covers costs for hackathons. The registration fee helped cover a fraction of the event costs. However, our biggest reason for charging a registration fee was to reduce the number of no-shows and to increase the likelihood that people would stay for the full day. We created a tiered pricing system:

- Students and Artists - $10
- Early Career Professional- $20
- Professional - $40
- Executive - $80

We allowed participants to select the category that was right for them and did not ask for proof of the category they chose. We made it clear in all of our materials that no one would be turned away for lack of funds. Everyone who requested a scholarship was invited to pay the student/artists rate if they could afford it and given the option to register for free.

Volunteers

During the planning process, we developed a list of volunteer tasks needed. These tasks ranged from set up and break down, staffing registration table, tech support, and time-keeping for the speakers and presentations. This list helped determine how many volunteers were needed. Many volunteers were staff from host and donor organizations. Some steering committee members were volunteers. Other volunteers were sourced from a general call to the arts community. In total, 20 people, in addition to the core leadership team and speakers, volunteered that day.

Other Key Decision Points

Other key decision points in the planning process included:

- Selecting the topic and articulating the goals of the event
- Identifying the intended participants
- Program
    - Length of time
    - Ensuring people are engaged throughout

    - Event date
    - Speakers
    - Location
- Budget and donors
- Datasets (readily available and wish list)
- Communications and marketing plan
    - Rollout
    - Logos and messaging
- Prizes
- Technical aspects
    - Website
    - Registration site
    - Dataset repository
    - On-site Internet connectivity (connectivity speed of over 100MPS recommended)

2. Compiling Datasets

Both the County and City of Los Angeles utilize an existing infrastructure to publish and share datasets through Socrata's Open Data platform. Other local governments and agencies are starting to publish raw data beyond what is presented in flat reports, but this is not always the case. No matter the topic of your datathon or data-driven event, think broadly about the issues that affect the ecology in which your topic exists. For the issue of arts access, we sought to be inclusive in how we defined arts and access. Therefore, we included datasets about library circulation, bookstores in LA County, public art, arts education, funding, demographics, and arts organizations.

The next step was to reach out to the authors or keepers of the datasets identified as relevant to our topic. This was a great opportunity to talk about the upcoming event and gauge interest and expertise on related fields. Many of our advisory committee members and some speakers were discovered during this process of compiling datasets. We discovered some nonprofits and private companies were willing to share aggregated datasets with us, as part of their outreach and support to the field.

Once we secured the right person to grant access or permission to the data, we discussed use. Often, the dataset we wanted was behind a paywall or needed some degree of cleaning and anonymization. We made the decision early on to only include datasets that the owners were willing to make available for free indefinitely,

so others who were not able to participate in the live event would be able to replicate or retrace the work done during the Arts Access Datathon. While some of the datasets were already licensed, we recommended the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License to all the owners of data previously unpublished.

We created several ways of presenting the data for all levels of participants. For the data novices, we created a narrative-based introduction to the datasets on the website. In this narrative we included a description of what the dataset contained and what kind of analysis each one might be useful for with its limitations. For those familiar with data in tabular (spreadsheet) form, we embedded an Airtable into a web page. With sorting and filtering features, the Airtable contains data download links and metadata fields to indicate the geography, periodicity, licensing, owner, related topics, and description of each dataset. Finally, for programmers and developers, we created a GitHub repository for select datasets (uploaded as csv files).

3. Designing the Contest

When thinking about the goals of the datathon, we discussed what possible deliverables could be produced by the project groups in the course of an eight-hour event. Since we expected the audience of the first Arts Access Datathon to primarily be made up of non-data people working in the arts, we veered from the usual hackathon format of awarding big (cash) prizes. Instead our goal was to inspire people to think about arts data, i.e. what they are, where they are, how they are collected, and how they can be used. We also reminded ourselves regularly that while the datathon was our tool, the primary focus of the day would always be on improving access to the arts.

Datathon participants were briefed on datasets and data tools through a series of five to ten minute talks and encouraged to spend time going through each of the resources together in their groups. In all, the participants had three hours to develop a project proposal that they would then present to a panel of judges. The panel of judges was carefully selected to mirror the diversity of the arts ecology. It included artists, arts administrators, and arts funders alongside technologists and data experts from outside the arts ecology.

To help the judges with their task, we provided them with a simple scoring rubric. They were asked to score each proposal on four key factors (the likelihood of improving access to the arts, use of data, feasibility, and creativity and innovation), with five points per factor. The judges did express some concern that they could not go back and change their scores after watching the first few presentations, but this

did not turn out to be an issue. There was general consensus from many after the event that the three winners were the best proposals.

There was some discussion within the leadership team about whether to award prizes or not. We decided that since our prizes were of nominal value, awarding them would not create undue incentives that might remove the focus from improving access to the arts. The prizes consisted of arts experiences (including public art tours, screenings, performance tickets), technical books, and gift certificates. Local arts organizations and tech book publishers donated most of the prizes.

4. Archiving and Preserving Event Materials, Media, and Datasets

Many data-rich resources, documents, and presentation files will be created during the planning and execution of any datathon. These may have historical value or use-value outside of the event itself. For example, the Arts Access Datathon website and presentations may be of interest to a variety of practitioners, researchers, and students interested in learning about or using open data. The goal of most archives is to provide access to users in perpetuity. In order to achieve this goal with digital files it is important to organize, contextualize, and securely store files as soon as possible.

Digital Archiving for Your Event

The most important aspect to remember when preserving and archiving digital resources is that they are not more stable or safer than physical resources (such as paper documents), and sometimes the opposite may be true. Technology changes at a rapid pace, and content is often lost from the internet or corrupted while stored on a hard drive or removable storage device. If your event generates media attention, social media activity, or maintains a significant web presence, it is also important to capture this while it still remains on the web. The following are some high-level considerations for digitally archiving a datathon event.

- **File Formats**
  Avoid proprietary file formats when possible, and choose formats that can be accessed by a variety of operating systems. This means that if a file requires a specific piece of software to open, you may be jeopardizing your ability to access the file at a later date. For example, while a particular piece of software may be favored for accessing or manipulating a file by most users today, this software might become obsolete in a few years. In addition, users with a different operating system may be unable to open it today.

  Recommended Formats statement from the Library of Congress:

https://www.loc.gov/preservation/resources/rfs/TOC.html

Resources for determining File Types:
https://www.archives.gov/preservation/products/definitions/filetypes.html

- **Naming Files**

  Consistent and detailed file naming can provide necessary context and organization, particularly when local knowledge and memory of the event and its corresponding materials have been lost or forgotten. Never use spaces or special characters in file names, and include a date. While some file metadata might keep track of "Last Opened" or "Created" date, these can change often or might not reflect the preferred creation date (for example when migrating to a different file format).

  Example of a bad file name:
  Photo 145.jpg

  Example of a good file name:
  0145_Judges_Datathon_04222017.jpg

  For the event, uploaded files to the GitHub repository utilized the following file naming convention: [Author/Owner-SubjectMatter-YYMMDDpulled].[file format]. For example, if the uploaded dataset is Los Angeles County Arts Commission's civic art collections data, the uploaded file was named: LACAC-CivicArtCollctn-170410.csv. As for abbreviations in the file name, these were left open so long as the abbreviation was explained in the Additional Documentation section.

  Additional tips for file naming are available from Stanford Research Library:
  https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming

- **LOCKSS**

  An important concept to keep in mind when preserving digital materials is "Lots of Copies Keeps Stuff Safe" (LOCKSS). The acronym originated from a peer-to-peer network of the same name. You might keep your files highly organized in the most stable formats, but if they exist solely on an aging external hard drive you might be out of luck sooner than you think. Making copies of files and keeping them in geographically dispersed places (including

external drives, local repositories, and in the “cloud”) can help ensure long term access to files when one storage site is compromised or fails.

"The cloud" refers to digital storage accessed remotely through the web. Common cloud storage platforms include Box, Dropbox, Google Drive, and iCloud. Each service maintains physical servers in different geographic locations and will have varying storage and security policies, as well as terms of use. No matter what service you use, it is important to keep local copies whenever possible and refrain from storing documents with confidential or private information on the cloud.

- **Web Archiving**

  The most well-known web archive is the Internet Archive and its Wayback Machine. The Internet Archive uses automatic crawlers that create shallow “snapshots” of websites and then make accessible archived versions available online. Even if a website is taken offline, you can still browse archived versions of the site by inputting the URL in the Wayback Machine and selecting a capture date.

  Not all websites are captured by the Wayback Machine automatically. As of May 2017, you can input an individual URL to be captured and included in the Wayback Machine using the “Save Page Now” feature.

  Other tools exist to create web archives that can be downloaded and preserved along with other digital assets from your event. The most robust and accessible tool is WebRecorder.io, a project of Rhizome/New Museum. Using their browser-based tool, you can record a website as you browse through it, and download a WARC file of the archive. WARC is the ISO standard for web archiving. Their Webrecorder Player can load WARC files from your computer offline to ensure the completeness of the archive.

  Web archiving is a not a backup of a website, and cannot be used to rebuild a website if it disappears. In addition, if a website utilizes complex Javascript or other dynamic features, it is possible that not all functionality will be captured.

- **Datasets**

  Datasets that your participants use will come from a variety of sources and are unlikely to be of the same file type. Perhaps the simplest and most stable format is CSV (comma separated values), which is ideal for moving data between tools. CSV files can be easily opened and manipulated by using

common spreadsheet software such as Excel, as well as reformatted into more functional formats such as GeoJSON.

Spreadsheets accessible online through such services as Google Docs and Airtable should also be downloaded as a CSV and stored offline.

The Arts Datathon website linked to various datasets hosted elsewhere, such as the County of Los Angeles Data Portal. In addition to these links, we also maintained a GitHub repository that stored data in spreadsheet form. This served as a snapshot of the data used at the event. In addition, it can serve as yet another copy of resources from the event for preservation purposes.

Datasets will come in a variety of formats in addition to CSV. To learn about the various formats, the Format Descriptions for Dataset Formats from the Library of Congress is an excellent resource: https://www.loc.gov/preservation/digital/formats/fdd/dataset_fdd.shtml

- **Create a Finding Aid or Data Manifest**
  Archivists will often create a Finding Aid that indexes or inventories objects (in this case, datasets) to assist researchers in accessing materials in a collection. While a formal Finding Aid may not be necessary, a manifest of the various items/files you are preserving along with minimal metadata including creators, copyright, and descriptions, can help maintain organization and institutional memory around an event.

  Automated Manifest creators can simplify the process when documenting hundreds or thousands of files. Karen's Directory Printer is a favorite among digital archivists.

For questions about the Arts Access Datathon, visit the website at artsdatathon.org, or drop us a line:

Bronwyn Mauldin
LA County Arts Commission
bmauldin@arts.lacounty.gov

Wendy Hsu, PhD
City of LA Dept. of Cultural Affairs
wendy.hsu@lacity.org