Organizing a Content Profile for a Large, Heterogeneous Collection of Interactive Projects

Eric Kaltman*, Roger Lorelli†, Adam Larson‡ and Ethan Wolfe§
Department of Computer Science, California State University Channel Islands
Camarillo, California
*eric.kaltman@csuci.edu, †roger.lorelli663@myci.csuci.edu, ‡adam.larson535@myci.csuci.edu, §ethan.wolfe734@myci.csuci.edu

***Abstract*—This paper details the organization of a "content profile" of a large longitudinal collection of interactive project prototypes of singular provenance. A content profile aims to analyze and summarize aggregate file metadata associated with a collection to aid in digital preservation strategies. Here, we detail the qualitative and quantitative methods used to organize a profile of a 14TB data set containing around 10.5 million files and 5,000 file extensions. The work extends the use of a content profile toward the historical characterization and interpretation of software development records. Additionally, the work prefigures further challenges associated with historical analysis of large, interdisciplinary data sets.**

***Index Terms*—computational archival science, computer games, software development, content analysis, history**

I. Introduction

The focus of many archival "big data" studies centers on large collections of text-based records that differ significantly in formats and organization. These so-called "heterogeneous" data collections provide significant challenges to large scale analysis due to issues ranging from file format conversion and text normalization, to data cleaning and data set derivation. When a lot of different formats are mixed together, significant work needs to be spent contextualizing their use and figuring out how to convert them into a generalized form of output without losing potentially salient information. However, in the case of interactive arts projects and software development records, the term "heterogenous" acquires a form of complexity. These projects seek to provide outputs that require multidisciplinary inputs. Art, text, video, 3D models, and other classes of records all combine into final, executable experiences. The variety of records present requires that we develop new strategies and methods for interpreting and organizing these collections.

This work aims to provide an overview of the issues and challenges associated with attempting to make sense of a large archive of prototypical interactive arts projects. It is a follow on analysis to earlier preliminary results obtained from a brief look at projects within a large, 18TB backup of prototypes from a leading interactive arts program [1]. That worked telegraphed a set of questions regarding the potential and challenges inherent to managing and interpreting heterogeneous collections of software-based records. Here, we present a further elaboration on that data set now that initial analysis of its entire scope has been completed. The goal is to better develop the methodologies behind organizing a "content profile" of a large data set. This involves using a variety of qualitative and quantitative methods to determine the data's context, provenance, and composition.

The work of "profiling" in this way is of benefit to multiple disciplines implicated in computational archival science (CAS) work [2]. Practically, organizing a data set makes it amenable to automated computational analysis; allows for the determination of issues relating to archival preservation and appraisal; and highlights how the data set might speak to various critical inquiries from humanist fields (in this case historical ones). The definition of a "content profile" is not common in the literature, but we use the basic definition from [3] as a starting point. Essentially, "profiling" involves aggregating the metadata associated with a collection of records into a unified representation that provides staging for further analysis. In the referenced work, the profiling sought to determine the implications of a data set's scope and content for archival concerns. There was a need to know more about the data's organization and form to improve its attendant data management strategies. Our work, while steeped in practical data management issues, seeks to broaden the application of profiling and look deeper into the various methods, both qualitative and quantitative, that became necessary for comprehending a highly varied, multi-terabyte collection.

This paper proceeds as follows. First, we highlight the necessary background from the preliminary analysis study, as well as related work describing the study of heterogeneous archival data sets. Included in this background as well is a brief summary of work involved with file format identification. We then proceed to detail both the qualitative and quantitative methods undertaken to organize a content profile of the project data into a unified database of derived file metadata. Following that, the results section highlights what was learned about the methodologies and tools necessary to conducting the profile. This leads to more elaborative discussion section about the implications of these methods as a means of archival and historical inquiry. We conclude with an elaboration on further plans and avenues for research into the data set.

978-1-6654-3902-2/21/$31.00 ©2021 IEEE

II. Background

A. Profiling Heterogeneous Data Sets

The categorization, profiling and exploration of heterogeneous document collections is nothing new to the field of digital preservation. While there does not appear to be a set definition of "heterogeneous" in the literature, the basic concept taken from [4] refers to "any data with high variability of data types and formats." Others refer to "heterogeneous" as synonymous with "messy" unstructured data collections [5]. In practice, heterogeneity has been applied to collections of singular provenance that mainly contain various forms of text document formats [6], web history collections [7], arbitrary collections of government working documents [8] and data sets aimed at identification tool validation [9]. [5] refers to their collection as "highly" heterogeneous. Given that our current data set is more diverse ("extremely" heterogeneous?) we believe that categorization and profiling of similar collections is relatively rare.

The noted applications for the heterogeneous data sets above mostly involve using them as test beds for the development of novel information retrieval algorithms. It may be that once a data set gets too diverse there is diminishing value in allowing the diversity to remain for analysis tasks. For example, [5] above used their highly heterogeneous data set to prototype a keyword extract tool for text-based documents in the collection, which was only subset of the file formats available. One major issue with dividing up a heterogeneous data set into more similar and manageable chunks is that certain contextual information might be lost or not recoverable without correlation with other content forms in the collection. The interactive projects in our data set are explicitly interdisciplinary, and feature a variety of programming, art, and project management records that may need to be analyzed on a project-level basis to afford historical and contextual insights. However, this does not preclude chunking the data set into like file types, it just notes that other work on heterogeneous data sets sometimes elides the heterogeneity that we are directly engaging with in this work.

B. File Format Profiling Identification

Our use of a "content profile" based on file format identification is borrowed from [3]. In the original context, the profiling is intended to provide "in-depth characterization [for] effective planning" of digital preservation strategies. As each form of digital data in a collection is tied to a specific file format, and that format is further dependent on a closed-set of software able to interpret it, the profile allows for a broad view of the preservation issues and constraints of a data set. Furthermore, the steps involved in said profiling, namely: "meta-data gathering, processing and aggregation, and meta-data analysis" roughly align with the quantitative analysis steps described both in the preliminary analysis and the Quantitative Methods below.

Crucially, the content profile work illuminates two concerns that we reiterate below. First, that the organization of a content profile for heterogeneous, unstructured, "real world" collections is too time and labor intensive to resource without further automation methods. And second, that inside highly heterogeneous collections, even determining what the more "homogeneous" classes of documents might be is difficult. What criteria do we apply to decide on format or data organization similarity? [3] provides an example of grouping documents by singular file formats, like PDF. However, if the data footprint of each PDF is highly variable then different methods will be required based on file size. In their example all three files are PDFs in known formats, but one is 120 times as large as the others. This shows a limitation of grouping by the homogeneity of file types alone without recourse to additional contextual and technical metadata.

The use and validation of file identification tools is common in digital preservation work. File identification is concerned with determining the correct historical format for a particular data object and, as indicated above, identification tools are commonly used in the analysis of complex data sets. There have even been recommendations to organize heterogeneous collections as a form of tool validation [9]. In this work, we make use primarily of the DROID File Identification Tool due to both its prominence in the literature ([10], [11], [12], [13]) and linkages with the PRONOM file format registry [2]. DROID analyzes a target file directory and provides organized and exportable reports that link each file type to a known file format (if possible). The tool correlates header information from a file as well as its file extension to identify potential file format types. Additionally, DROID can be configured to provide unique hashes for each file encountered.

A limitation of the tool, however, is that the PRONOM database is not an exhaustive listing of file format information. The total registry has around 2300 formats identified[1], and there are some issues with file disambiguation. This mainly affects formats associated with long running production software suites, notable the Adobe Creative Suite, where header information alone is sometimes not enough to determine proper versioning. Regardless, the DROID tool is configurable and available for distributed command line use, which makes it easy to integrate into scripts. DROID's speed is also a bit of an issue, and others have used high performance computing to get around its limitations for analyzing large data sets [14]. There are many other potential file identification tools and we are looking into the use of a larger suite of them to compare with this initial content profile.

Another project that is methodological similar to our quantitative work is that of the Brunhilde Tool [15] created by the Canadian Centre for Architecture (CCA). Brunhilde makes use of the Siegfried file identification tool [16] to characterize the file metadata for individual directories. It accomplishes this by converting the reporting output of Siegfried into a queriable SQLite3 database. The tool then uses that database to generate

[1]The current DROID XML signature file lists 2246 formats as of October 2020, see: https://cdn.nationalarchives.gov.uk/documents/DROID_SignatureFile_V97.xml

an HTML report including salient statistics and summaries of the files. Brunhilde's reports are similar to the proposed content profiling work above and are also aimed at archival management concerns. It should be noted that the goal of this work is also historical investigation, that is the scholarly use of historical heterogeneous data sets, in addition to archival concerns. We believe the overlap in methodologies portents a more thorough future alignment of archival and historical analysis tools for big data sets.

III. Methodology

The data set constitutes 18TB and contains the working documentary backups for 546 interactive project prototypes developed from 2001-2019 by Masters students in a leading interactive arts program. Projects were completed in 14-week, semester long development "sprints" by diverse, interdisciplinary teams that paired faculty advisors (all with professional industry experience) with client organizations. The projects ranged from computer games, novel interface designs and entertainment systems to museum installations and computer animation projects. Teams were organized to allow for students to collaborate across technical and disciplinary boundaries, and constrained by the professional requirements of clients. The work produced in the program is akin to the prototyping, pre-production processes one might encounter in the entertainment industry, and the program's graduates generally go on to work for leading companies in their respective entertainment sectors. It should be noted, that as these works are prototypes designed on short schedules their documentary organization is probably not consistent with what one would find in larger production projects completed along normal industry timelines.

The ability to provide a large-scale analysis of interactive (and therefore computational) projects is rare due to numerous legal and technical issues. Access to production data is not commonly shared for fear of revealing the trade secrets, interpersonal dynamics and "mess" of production work [17]. Additionally, the value of these records is not considered a priority within the industry, focused as it is on entertainment outputs and their consumption by users [18].

A. Notes from a Preliminary Analysis

An initial preliminary analysis of the data set investigated a small subset of the projects (three computer game and one animation) from the Spring 2006 semester [1]. The four projects were originally chosen on the misguided assumption that earlier project work would have less data to analyze. This proved incorrect, with the animation project alone accounting for a majority of data by both file count (109,683 out of 142,059) and data size (275GB out of 293GB). Analysis involved a brief look into the content of each project's files to determine a project's purpose and goals. This was followed by a file format analysis in which the DROID file Identification tool [19] was used to output file-level metadata into a unified database that allowed for basic metadata querying. The processes described in both the qualitative and quantitative methodologies below are significant expansions of this initial work, both in the scope of content analyzed and the scale of files identified.

The preliminary work noted that the content of this data set, as a longitudinal collection of single-origin provenance, could answer a number of research questions in history and archival science related to the chronology of development software packages and environments as well as the general content form and access constraints for historical production collections. Additionally, the preliminary work pointed toward the potential for further computational methods to be used once the data set was characterized. This potential is elaborated on in Future Work below.

B. Transitioning Toward a Full Content Profile

A team consisting of the primary investigator and three undergraduate research associates in the Software History Futures and Technologies (SHFT) group spent two months modifying and expanding the methods from the preliminary analysis to cover the entirety of the 18TB data set. The full data set consisted of around 10.5 million files and 5000 file formats produced over a 20-year period. During initial review, it was decided to remove 4TB of academic course-related data from the analysis as its structure was qualitatively different from the project semester organization found in the preliminary analysis. This left 14TB for organization into a full content profile. The analysis proceeded in two general phases. In the first phase, a primarily qualitative process was used to organize derived, project-level metadata about each project. This was followed by a second phase that linked the initial qualitative results to a quantitative analysis of the file-level metadata inside every identified root project folder. Due to the scale and complexity of this data set, a complete content profile is still forthcoming, as it will likely require further publications to completely explicate and contextualize.

The data set was initially distributed to researchers in the form of four 10TB external drives that arbitrary broke up the initial, unified backup data set. Many of the drives were not full, and the first step was to duplicate the 18TB data set twice between the four 10TB drives. After that, a third, unified backup of the data was copied to a shared network attached storage (NAS) device. This "canonical" store was then used to further distribute the data set to three additional 8TB drives for local use by research associates. [2]

C. Qualitative Organization of Contextual Information

Each of the projects originated from a singular provenance, which allowed for a consistent initial qualitative study of the data set. The academic program itself organized each project around a set of four team milestones throughout a given semester. As a result, many projects included summary

[2]This data distribution scheme was necessary due to the influence of the COVID-19 pandemic, as the research associates needed to use their home machines (or university provided ones) that did not have the necessary storage space to hold the entire data set. Local bandwidth limitations and caps prevented easy network-based data sharing, so local, wired USB 3.0 access was used.

materials oriented around quarter, half, three quarter and final semester project reviews. Additionally, the program provided summaries and links to each historical project's development website through a unified past projects page on the program's website. Lastly, during the later half of the 2000's the program initiated an "Archival Bits" documentation strategy to try to better preserve project closing information. The program provided a default project folder organization scheme and checklist for final project submissions. This last strategy was followed haphazardly but still enabled easier analysis for the projects that made use of it.

The projects' organization on disk is loosely sorted by project chronology. Projects developed before 2006 are lumped together in a general directory, whereas post-2006 each project resides within a containing folder that labels the year and semester of development. Some projects were also found nested within each other due to continuing development across multiple semesters or interdependence of project subcomponents. Each project team was responsible for their project folder, so the organization varies significantly from project to project.

The first step of the qualitative analysis involved organizing the information provided by the program's website into a spreadsheet listing each known project by name and date. This was then expanded to include a collection of 15 fields noted in Table I. The fields relating to technologies used and file path information required the most exhaustive qualitative work. Each research associate received a subset of the full data set, with the target project folders tracked on a separate spreadsheet. The work proceeded project folder by project folder. Processes for locating files and determining project characteristics varied by associate, as the variety of projects and file organization was not amenable to any standardized initial process. This does not mean that certain methodologies did not arise from the collaborative investigation. During the later stages of the work certain commonalities were derived and support tools enlisted to speed up the characterization work. For example, in order to locate project files, the associates used the WizTree Disk Space Analyzer [20] to directly search each volume's Master File Table (MFT). Searching the MFT allowed them to perform quick keyword searches for known variations of project names within the data set.

In addition to determining the creators of each project, special notes were made of the underlying technologies used. Many of the systems and infrastructures employed are no longer in use, having been either deprecated within the entertainment industry, replaced by newer versions, or never released. Determination of the technical context for each project is necessary to better understand the file structure and organization within a project, as well as the secondary applications that generated project files. There are many forms of intermediary production involved with interactive projects, and much of their historical context is still unknown due the relative lack of access to similar collections.

D. Quantitative Organization of Data Set Structure and Content

The quantitative organization of the data set's file metadata followed a similar, though much expanded, process from the preliminary analysis. In that work, we characterized four projects by running a top-level directory scan with the DROID analysis tool, and then inserted the DROID CSV report output into an SQLite3 database. The script managing the conversion process was written in Python 3.0 in a JupyterLab notebook. That notebook formed the basis for the expanded characterization of the full data set. While the process pipeline, DROID to CSV to relational table row, was not complex, running the analysis on the entire multi-terabyte data set did cause some issues.

Because the scope of some individual projects ranged into the hundreds of gigabytes, numerous top-level semester folders contained over a terabyte of data. Due to space and memory constraints on the associate's home machines, it was not feasible to run the analysis scripts locally. Fortunately, the research team had access to the SHFT group's local dedicated server. A full copy of the data set was transferred to two 10TB drives and mounted on the dedicated server. The team then abstracted the initial process script from JupyterLab into a small command-line tool that initiated DROID's command line interface and generated DROID CSV reports from a listing of input directories. The command-line tool spawned multiple DROID processes that themselves ran multiple threads across two VMs (as each 10TB drive was locally mounted in a different VM). Even with this setup, the initial metadata aggregation through DROID took more than a week alone. This is likely due to bandwidth limitations involved in concurrent access to a single USB 3.0 interface.

After the generation of the 546 individual project reports as DROID CSVs, the CSV data was stored in the group's shared cloud storage folder. This allowed the CSVs to be downloaded in bulk and another script was written to insert the CSV data into a local SQLite3 database. Once generated, this SQLite3 database was then shared between associates for preliminary analysis.

IV. Results

A. Exploratory Analysis Methodologies

The qualitative results highlight two larger routes of study. First, the manual characterization of each of the project-level data in the project spreadsheet allowed for a basic historical contextualization of the program's work over the last twenty years. This enabled a rudimentary look at the changes in development environments and technologies, as well as categorization of subsets of the data set for further analysis. Based on the qualitative categories, we found that a majority of the projects were partially developed in either the Unity Engine (39%) or Adobe Flash (16%) environments. Other prominent project categories included computer games (56%), mobile applications (14%), physical installations (14%), virtual reality

TABLE I
QUALITATIVE ORGANIZATION RAW INFORMATION FIELDS

Field	Description	Data Form	Reason for Inclusion
Project Name	Public name of the project[4]	Text	
Category	Description tags used to categorize projects for further subset studies	Multiple tags	This field made use of an informal taxonomy to allow the project team to tag different development paradigms and application types. For instance, a project might be tagged as "Unity" and "VR" to help quickly locate those types of project for further studies.
Description	A formal description of the project copied from the program's public website	Text	Provided an initial contextualization for the project
Project URL	A link to the archived project website	URL	Each project, by program mandate, had a public website to provide information on project development and outputs.
Semester	Semester that the project took place, numbered 1, 2, and 3 for "Spring", "Summer" and "Fall"	Text	Chronological sorting
Year	Year of project work	YYYY	Chronological sorting
Project Advisors	Program advisors assisting with the project	Text	Correlation for future interviews and comparison of organizational schemes
Project Members	Student project members	Text	Correlation for future interviews and comparison of organizational schemes
Client	Organizational entity responsible for initial project requirements	Text	Correlation for future interviews and comparison of organizational schemes
Technologies	Development technologies and dependencies used on the project (broadly)	Text	Categorization for broad dependencies to allow for future grouping by application architecture and software dependencies
Project Data Online?	Project data in some cases was also posted to public web resources and differed from data located in the backups	Yes / No	Indicated need for future web scraping of content
Parent File Path	Normalized path to the root directory for the particular named project	UNIX directory path	Root directory from which to originate DROID analysis scan[5].
Executable File Path	Normalized path(s) to executable file locations	UNIX file path	Locating possible targets for emulated execution (or native execution if recent)
Documentation File Path	Normalized path(s) to project documentation directories	UNIX file path	Locating possible documentation for project outputs and organization

(10%), and computer animation (7%). As expected, the transition away from Flash-based applications occurred around the turn of the 2010s. Only 12 projects made use of Flash after the Fall 2010 semester.

Second, the qualitative analysis process itself is ripe for reflexive critique. In this mode, the means by which the associates organized the qualitative information becomes the object of study. While not apparent at the outset of the project, it became clear that the associates developed a particular set of preliminary strategies when approaching a new, un-analyzed project folder. These approaches were initially informed by the historical context of the program's management and course structure. In discussing the qualitative work above, we mention that the singular provenance of the projects allowed for certain base assumptions about the organization of project files, and crucially, where to go first to characterize and summarize the project. Projects of a specific categorization prefigured a set of coordinated practices by the associates, who found investigating computer game-based projects easier than other forms of production work. This is likely due all the associates having experience with modern game development work as the general structure of game projects, delineated as they are into various source code and asset directories, has not changed much over the last twenty years.[3]

[3]We note that the projects in this collection are an example of advanced student productions. Perhaps the organization of game projects at the semester-scale has not actually changed. We cannot make definitive claims about professional development work trends based on this data set.

The categorization scheme is currently not controlled nor is it mutually exclusive. Due to the pioneering work of the program, many projects included the coordination of numerous development environments. For example, the Voyage project from Fall 2017 [21] was a VR-based classroom experience developed in Unity and managed with a separate Apple iOS program on an iPad. This shows that modern development work involves many secondary software dependencies as well as overlapping technological categories. It would be a disservice to future historical and archival work to attempt a singular classification for these projects. It may be beneficial to develop a more robust taxonomy of digital project classifications, as many other realms of cultural heritage production (like digital humanities) also work with mixed methods and media. Certainly, more delineated project-level classifications would help historians prepare specific technical methodologies for different project types.

B. Preliminary Data Set Content Profile

Based on the initial content profile generation described above, the data set contained 9,202,496 files (excluding directories) with 4,982 distinct extensions and occupied 14.63TB of total storage space. Due to limitations of DROID file identification, 3,425 extensions were not characterized (totalling 4,042,379 files and 44% of the data set). Additionally, 1,144,294 files (22%) were only characterized by file extension alone, which is a less conclusive identification than ones making use of file content and header information [13]. The resulting file metadata database was 6GB in size, containing two tables with 21,073,510 rows. Initial data set file format statistics are summarized in Tables I, II, III.

Of note is that 36% of the data set was redundant. DROID allows for a hashing flag to be set that generates a unique hash id for each file. Using this hashing, we were able to check data duplication for the entire data set. Much of the redundant files are either duplicated project assets, like images, or duplicated development environment files stored over multiple projects. In many cases, particularly in Unity Engine or Adobe Flash environments, certain core library and development files area large portion of standard deployment packages.

Similar to the preliminary analysis, a large amount of the files remain unidentified due to their place as intermediary production outputs. As shown in Table IV, .ress, .mp4 and files with no extension account for 918GB total storage space in the data set. In total, unidentified files accounted for 12% of the total data set by size.

Of the formats that were identified, video-based files occupied the most storage space. As seen in Table III, M2TS high-definition, Quicktime, and AVI videos took up 6.7TB (46%) of the total data set. In looking at Table II, the most files of a single format are generic .bin files. The content of these files is currently unknown, so they are identified here only in the strictest sense, that of a file extension match with an existing PRONOM format. Of note is that the specific PRONOM entry for .bin files is explicitly under defined. This leaves .png with the highest number of known identified instances with 869,223 files.

Generating the profile database also encountered some unforeseen edge cases with the DROID tool. In one memorable instance, the tool was unable to characterize a group of files with a common distribution directory of the Node.js web application framework. It turned out that those specific files were themselves created as edge case tests for a file parsing tool, and as such, included program breaking modifications like complex, overlong file names, odd character and encoding configurations, and intentionally corrupted file metadata. This was an instance of a historical set of test cases testing our historical analysis test case.

TABLE II
FILE FORMATS BY NUMBER

Extension	File Format	Format Version	Count
bin	Binary File		618453
png	Portable Network Graphics	1.0	587617
	Portable Network Graphics[6]	1.0	281606
jpg	JPEG File Interchange Format	1.01	233276
xml	Extensible Markup Language	1.0	228147
jpg	JPEG File Interchange Format	1.02	184594
tif	Tagged Image File Format		182537
php	Extensible Hypertext Markup Language	1.0	180855
png	Portable Network Graphics	1.1	176310
html	Extensible Hypertext Markup Language	1.0	146563
class	Java Compiled Object Code		129911
js	JavaScript file		114332
exr	OpenEXR	2	106891
wav	Waveform Audio (PCMWAVEFORMAT)		105620
dll	Windows Portable Executable	32 bit	92142
png	Portable Network Graphics	1.2	87158
info	Guymager Acquisition Info File		76318
iff	Maya IFF Image File		74963
py	Python Script File		63072
jpg	Exchangeable Image File Format (Compressed)	2.2.1	60504

[4]A few projects had different internal project and folder names from their public description on websites. This led to difficulties in identifying certain projects.

[5]In some cases, there were multiple potential root folders, or root folders were nested inside other project's roots.

[6]This row lists pngs identifed without a proper file extension

TABLE III
FILE FORMATS BY AGGREGATE SIZE

File Format Name	Version	Size (GB)
M2TS		2428.36
Quicktime		2394.24
Audio/Video Interleaved Format		1892.8
MPEG-4 Media File		872.33
Adobe Photoshop		501.28
Tagged Image File Format		417.65
Maya ASCII File Format		397.33
Portable Network Graphics	1.0	340.53
OpenEXR	2	322.61
Macromedia FLV	1	187.17
Exchangeable Image File Format (Compressed)	2.2.1	154.24
Maya IFF Image File		133.54
Microsoft Powerpoint for Windows	Multiple	121.23
JPEG File Interchange Format	1.02	114.21
Waveform Audio (PCMWAVEFORMAT)		104.1
Raw JPEG Stream		103.19
Java Compiled Object Code		98.9
Windows Portable Executable	32 bit	96.39
MPEG-2 Program Stream		95.43
Apple Lossless Audio Codec		92.36

TABLE IV
UNIDENTIFIED FILE EXTENSION BY AGGREGATE SIZE

File Extension	File Count	Size (GB)
ress	5401	375
mp4	2659	310
	1216602	228
assets	15101	196
abc	479	80
mov	649	53
pdc	58074	34
cfa	1351	31
mc	5504	30
obj	17926	28
resource	12689	27
avi	64	24
bnk	516	21
vmdk	14	21
xnb	18870	18
uasset	8696	15
egg	22037	13
m4v	130	12
pack	340	12
unity3d	259	11

V. Discussion

A. Limitations of File Format Identification

As briefly noted in the preliminary profile above, a large number of the file formats in the data set could not be identified automatically. Some of these, like certain Adobe Flash and Autodesk Maya formats already appeared in the preliminary analysis work two years ago. While they still remain unidentified, there are now 3,423 additional extensions that could not be linked to a PRONOM entry. This might, however, be due to the limitations of the DROID tool itself and the sparsity of PRONOM coverage for more esoteric formats. The literature indicates that Siegfried analysis might produce more identifications [15], and we are now planning to run the same profiling process again with a different suite of file identification tools.

Regardless, these results do indicate that production process file formats, and other items that are part of intermediary production tasks [17] might not be well represented in file format registries. There is a distinct lack of support in PRONOM for a number of common software development file extensions, most notably those for the C/C++ programming language (.c, .cpp, and .h). Further, specific development environment files, like those for the Unity Engine (.meta) also lack entries. As the PRONOM registry is organized by voluntary contributions to the registry, it may just be that software development collections are relatively uncommon in archives and at other institutions making use of file identification tools.

B. Qualification of Exploratory Methods

The variety of strategies employed to contextualize the project's data, and by broader extension their different categories, aligns with the rising focus in data science and computing history on "data cultures" [22] and the "hermenetics" of historical data management [23]. Each sub-culture develops a set of strategies for managing, producing, and sharing data. In most cases, the discussion focuses on research data sets and scholarly communication. This work points to a broader context for data cultures in the interactive arts, and for the need to develop investigative strategies tied to each culture's historical data production methods.

VI. Future Work

The next phase of this work will proceed along a number of different research vectors. As the entire data set is now characterized along the qualitative and quantitative lines described above, work will now proceed to fully analyze the resultant metadata database and break up the qualitative projects into sub-categorizations for more granular analysis. The metadata database will allow us to answer a number of questions regarding the historical chronology of development files, periods of development activity, and delve deeper into unidentified file formats discovered. We are also planning on modifying our initial workflow to be more automated and to make use of other file identification tools, most notably Siegfried, to allow for comparison with the DROID results.

This analysis has also already spawned research efforts into better characterization of the Unity and Adobe Flash projects within the data set. Each represented a significant portion of the overall project development environments, and we are looking into designing automated tools for generating project metadata and dependency information for such projects. Additionally, the amount of data duplication in the collection inspired the creation of a tool for automatic data reduction that will preserve links to identical file references across the data set and re-encode and compress video and image assets.

Finally, based on the sheer scale of the data set, organization for a variety of machine learning methodologies might be practical and fruitful. A significant amount of qualitative work involved simply locating specific projects and technologies with the data set. It may be possible to train supervised approaches on both the qualitative categorization (through a more structured set of classifications and features) and the quantitative organization (by analyzing file content, quantity, and relative directory structures) to enable automatic tagging and search based on project type and dependent technologies.

VII. Conclusion

The scale of this data set (14TB and 10.5 million files) and its heterogeneity (5000 file extensions) points to a growing set of concerns for the historical study and archival organization of similar collections. Software development and other forms of digital production work create a voluminous amount of data and are generally unavailable for large scale, dedicated analysis of the kind found in this work. Furthermore, the data set analyzed above contains a significant amount of files that may not turn out to be historically salient or important. Without better understanding the historical context and organization of development projects, or for that matter, any modern data set derived from multi-disciplinary production processes, it will become continually more difficult to appraise, store, and access them. By detailing both the qualitative and quantitative work involved in the content profile of this data set, it is hoped that more attention to and delineation of historical data set contextualization processes can result.

Acknowledgment

This work was supported, in part, by California State University Channel Islands through both Faculty Mini-Grant and Summer Undergraduate Research Fellowship (SURF) awards. The authors would also like to thank the Entertainment Technology Center at Carnegie Mellon University for access to their backup collections.

References

[1] E. Kaltman, "Preliminary Analysis of a Large-Scale Digital Entertainment Development Archive: A Case Study of the Entertainment Technology Center's Projects," in *Proceedings of the 2019 IEEE International Conference on Big Data*. Los Angeles, CA: IEEE, Dec. 2019.

[2] T. N. Archives, "PRONOM | Welcome," publisher: The National Archives, Kew, Surrey TW9 4DU. [Online]. Available: https://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[3] P. Petrov and C. Becker, "Large-scale content profiling for preservation analysis," in *Proceedings of the 9th International Conference on Preservation of Digital Objects*. Toronto Ontario: iPRES, Oct. 2012.

[4] L. Wang, "Heterogeneous Data and Big Data Analytics," *Automatic Control and Information Sciences*, vol. 3, no. 1, pp. 8–15, Oct. 2017, number: 1 Publisher: Science and Education Publishing. [Online]. Available: http://pubs.sciepub.com/acis/3/1/3/index.html

[5] A. S. Maiya, J. P. Thompson, F. Loaiza-Lemos, and R. M. Rolfe, "Exploratory analysis of highly heterogeneous document collections," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '13. New York, NY, USA: Association for Computing Machinery, Aug. 2013, pp. 1375–1383. [Online]. Available: http://doi.org/10.1145/2487575.2488195

[6] N. S. Alamri and W. H. Allen, "A comparative study of file-type identification techniques," in *SoutheastCon 2015*. Fort Lauderdale, FL: IEEE, Apr. 2015, pp. 1–5, iSSN: 1558-058X. [Online]. Available: https://ieeexplore-ieee-org.ezproxy.csuci.edu/document/7132993

[7] S. Peroni, F. Tomasi, and F. Vitali, "The aggregation of heterogeneous metadata in web-based cultural heritage collections: a case study," *International Journal of Web Engineering and Technology*, vol. 8, no. 4, pp. 412–432, Jan. 2013, publisher: Inderscience Publishers. [Online]. Available: https://www.inderscienceonline.com/doi/abs/10.1504/IJWET.2013.059107

[8] G. Rehm, M. Lee, J. Moreno-Schneider, and P. Bourgonje, "Curation Technologies for Cultural Heritage Archives: Analysing and transforming a heterogeneous data set into an interactive curation workbench," in *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, ser. DATeCH2019. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 117–122. [Online]. Available: https://doi.org/10.1145/3322905.3322909

[9] A. Fetherston and T. Gollins, "Towards the Development of a Test Corpus of Digital Objects for the Evaluation of File Format Identification Tools and Signatures," *International Journal of Digital Curation*, vol. 7, no. 1, pp. 16–26, Mar. 2012. [Online]. Available: http://www.ijdc.net/index.php/ijdc/article/view/201

[10] W. Underwood, "Extensions of the UNIX file command and magic file for file type identification," *Georgia Tech Institute of Technology*, 2009, publisher: Citeseer.

[11] L. Shala and A. Shala, "File Formats - Characterization and Validation," *IFAC-PapersOnLine*, vol. 49, no. 29, pp. 253–258, Jan. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896316324880

[12] A. N. Jackson, "Formats over time: Exploring uk web history," *arXiv preprint arXiv:1210.1714*, Oct. 2012.

[13] D.-G. Akestoridis and T. Henderson, "Poster: Enabling useful data sharing through format identification and text mining," 2016.

[14] R. Arora, M. Esteva, and J. Trelogan, "Leveraging High Performance Computing for Managing Large and Evolving Data Collections," *International Journal of Digital Curation*, vol. 9, no. 2, pp. 17–27, Oct. 2014, number: 2. [Online]. Available: http://ijdc.net/index.php/ijdc/article/view/9.2.17

[15] T. Walsh, "Data-Driven Reporting and Processing of Digital Archives with Brunnhilde," *Practical Technology for Archives*, no. 8, Jul. 2017, https://hdl.handle.net/1813/76867.

[16] (2021) Siegfried. [Online]. Available: https://www.itforarchivists.com/siegfried

[17] E. Kaltman, "Attending to Process and Data," *ROMchip*, vol. 2, no. 2, Dec. 2020, number: 2. [Online]. Available: https://romchip.org/index.php/romchip-journal/article/view/117

[18] J. Scott, "Saving Game History Forever - Or Dooming It To Oblivion?" San Francisco, CA, Mar. 2015, library Catalog: www.gdcvault.com. [Online]. Available: https://www.gdcvault.com/play/1022240/Saving-Game-History-Forever-Or

[19] T. N. Archives, "File Profiling Tool (DROID)," last Modified: 2017-09-04 Publisher: The National Archives. [Online]. Available: http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/

[20] (2021) Wiztree disk analyzer. [Online]. Available: https://www.diskanalyzer.com/

[21] (2017) Voyage. [Online]. Available: https://www.etc.cmu.edu/projects/voyage/

[22] C. L. Borgman, "The conundrum of sharing research data," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 6, pp. 1059–1078, 2012, _eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22634. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22634

[23] A. Acker, "Toward a Hermeneutics of Data," *IEEE Annals of the History of Computing*, vol. 37, no. 3, pp. 70–75, Jul. 2015, conference Name: IEEE Annals of the History of Computing.