Preliminary Analysis of a Large-Scale Digital Entertainment Development Archive: A Case Study of the Entertainment Technology Center's Projects

Eric Kaltman
Computer Science Department
California State University, Channel Islands
Camarillo, CA, United States
eric.kaltman@csuci.edu

***Abstract*—This paper describes a research plan for the investigation of the project archive from Carnegie Mellon University's Entertainment Technology Center, an interdisciplinary professional Masters program in interactive entertainment and game design. Representing nearly 20 years of project design in the interactive arts, the ETC's ad hoc archive provides a prospective template collection for the analysis of entertainment software projects through historical, archival, and computational methods. The work-in-progress described here is based on a preliminary analysis of four early projects from the archive, and provides a guide to potential synchronic and diachronic investigations of software development methodology in a chronological collection of unified provenance. Access to software process documentation is difficult to come by and this "working" collection provides a significant resource for approximating the organizational state of future software development collections to be ingested into archives.**

***Keywords*—computational archival science, computer games, software development, documentation, history**

I. Introduction

Access to development records is a significant impediment to the historical investigation of computer games and interactive entertainment. Due to market considerations, including non-disclose agreements and trade secrets, the physical and — now primarily — born-digital artifacts of software development processes are unavailable to historians of science and technology, and further, any digital humanist scholar wanting access to documentation on modern creative processes. While there is an extensive amount of software documentation and open-source code available on the Internet, full project documentation including all assets involved in production and iterative design phases is usually unavailable.

The Entertainment Technology Center (ETC) at Carnegie Mellon University (CMU) is a premier graduate interdisciplinary program in creative technology and art practice cofounded out of CMU's College of Fine Arts and the School of Computer Science. ETC focuses primarily on semester-long, client-based projects dealing with new and innovative technologies for game-based, mixed-media entertainment and amusement-based productions. In recent times, this has included significant work in emerging areas like augmented reality (AR) and virtual reality (VR). Since ETC's founding in 2000, they have aligned their work with emerging technologies due to their dedication towards preparing students for work in the film, game, and amusement industries. Most pertinent to the goals of computational archival science (CAS), ETC has also collected and preserved, in an ad hoc archival structure, the total intermediary and final documentary production for all of their projects. (Hereafter referred to as the ETC archive.) It is estimated that the total data footprint of the project folders alone, excluding their source code and closing documentation repositories, exceeds 30 terabytes.

This paper provides a background to the ETC archive's organization, and a road map, informed by previous work with game development documentation, for the investigation of ETC's data and its implications for our understanding of the last 20 years of innovation in the interdisciplinary media arts. Below, we elaborate on the ETC, its project data, and our experiences with game and software development documentation. We then proceed to an initial set of research questions in historical inquiry, archival science, and computer science, informed by a simple content analysis of some select ETC projects.

Due to the size and scale of the archive, even organizing the initial data set for analysis has proved challenging, as a result, this paper represents a work-in-progress and research plan for this material.

II. Background

This section will provide contextual background information on the ETC archive and an outline of the current state of game and entertainment software project archiving.

A. The ETC Archive's Organization and Provenance

The ETC prides itself on a consistent and rigorous curriculum based around diverse, interdisciplinary project teams working at the forefront of current entertainment technologies. ETC's faculty are composed primarily of industry veterans from the computer game, film, and amusement (theme-park and immersive technology) industries. Drew Davidson, ETC's

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

current director, sums up the ETC's process as a form of 'creative chaos', allowing just enough freedom to innovate while also functioning within industry-minded development constraints. ETC's two-year Masters in Entertainment Technology (MET) provides students with an initial initiation semester into ETC's creative outlook followed by three semesters of 14-week project development sprints. Each development team is intentionally organized along similar lines to ETC's admissions categories, in which 40% of students are technical (programming, software and hardware development), 40% are artistic (UI / UX, digital arts and music), and 20% are from a variety of creative disciplines (creative writing, fine art, etc.). Projects are categorized according to three general areas [1]:

1) **Transformational games** "for impact" in education, health, and other "meaningful gameplay" applications.
2) **Innovation by design** to further development of "next-gen interfaces, experimental prototyping,...[and] creative development practices"
3) **Interactive storytelling** "to design the most engaging narrative experiences [in] location-based [activities], mobile, transmedia, augmented reality and live interactive performance."

Each category is oriented around iterative development practices and a belief in creative failure as a means of growth. As a result, most of the projects in the archive represent — at that time — cutting edge research and development in the media arts.

An ETC project is conceived in one of three contexts: client-based work with an external partner, internal support for an ETC research project, or from a student group proposal. The 14-week development cycle features quarter and half semester check-ins to solicit wider feedback. The check-ins consist of descriptive videos, demos, presentations and project progress documentation. Final presentations and deliverables are required for the conclusion of projects, and there is an internal archiving process whereby project teams are asked to clean up project folder documentation and submit their data to the backup server and on physical hard drives.

The collection available from ETC is therefore unique in the level of contextual documentation provided for each project, and in that each project is conducted according to similar time and team constraints as well as development goals. In addition to the contextual documentation, all project promotional websites are saved and all code produced is stored in its final configuration in a project folder along with its version-controlled backup in ETC's Perforce version control server. While the additional documentation might be made available in the future, the work presented and telegraphed in this paper is based exclusively on access to the team project directories. Due to the size and scale of the collection, which include around 20 years of two to three semester project cycles, full analysis of the archive's 30TB+ footprint will take some time. For the purposes of this paper, and due to a time limitation based on data transfer logistics, we will only briefly look at the composition of four projects from the Spring semester of 2006. These are the earliest projects organized by date in the archive.

The potential for the larger collection should be possible to extrapolate from the initial project group given that all projects are of well-documented and similar provenance. The consistency of the project's cycles and timelines also allows for nuances in analysis that would not be possible for more general, multi-institutional or multi-user, collections of development data. Specifically, certain questions of historical progression and comparison become possible given that this is a chronological collection of 37 semesters of 5-10 projects plus an additional hundred projects not currently aligned with a particular semester.

B. Game Development Documentation and Archives

Full project software development documentation is difficult to find and analyze. This is mostly due to the commercial nature of significant development projects where competitive market imperatives prevent access to proprietary documentation. The projects available to analyze are, therefore, mainly online, open-source, and non-commercial. As a result, there is little available information and analysis of commercial project documentation, including little access to commercial source code and project organization methodologies. Below, we briefly outline some known collections of materials before turning to our previous work in archival game development documentation analysis.

1) Development Archives: Major repositories of game and entertainment project documentation are not common in institutional archives. Aside from the UC Santa Cruz and Stanford University documentation mentioned in part III below, remaining significant collections of game development projects are limited to further collections at Stanford, particularly game designer Hal Barwood's of LucasArts, game designer Warren Spector's archive at the University of Texas, Austin [2], and the collections of the Strong Museum of Play in Rochester, NY [3]. Most of Strong's development documentation is also contained in the collected papers of famous game designers. Coordinated ingestion of technical project documentation, even in science and technology disciplines, is relatively limited at this point. The situation also extends to digital arts and media collections as well.

There are known significant commercial game development archives, including collections at Nintendo, Electronic Arts, Activision-Blizzard, Microsoft, Nexon, and Disney Entertainment, but access is not open to researchers and certainly not to documentation in a similar intermediary state to much of the ETC archive. Use and analysis of the ETC data, especially if it appears to be commercially beneficial, may help to convince industry actors to allow for more research access to their development documentation.

2) Closing Kits: The games industry (and general software development industry) do encourage practices in project closing documentation. Heather Chandler, in [4], describes the process of a "closing kit" in which all final documentation, code, assets and other potentially future usable project products are recorded and stored. However, the closing process is not an archival process so much as a backup and storage one. It is also unlikely that a closing kit would include some of the intermediary information present in the ETC projects. ETC does, as mentioned, include a closing procedure for all projects according to an archival checklist which explicitly excludes intermediary documentation, instead working to collect project final presentations, finalized code, and other compiled outputs. Work with the ETC archive could shed light on what levels of closing documentation are appropriate and useful to future developers.

III. Previous Work

The methodology described in this paper is based on previous work by the author in the description and analysis of game development documentation. Two projects are highlighted, one for the National Endowment of the Humanities (NEH) aimed at academic research game productions (a very similar use case to some ETC projects), and an archival treatment of game designer Steve Meretzky's game development records in Stanford University Libraries Special Collections.

A. A Unified Approach to Preserving Cultural Software

Our NEH grant, "A Unified Approach to Preserving Cultural Software Objects and Their Development Histories", was a joint project between UC Santa Cruz Library, UC Santa Cruz's Computational Media Department, and Stanford University Libraries looking into the appraisal and description issues inherent to academically produced research computer games [5]. In this case, the game in question was *Prom Week*, a social simulation of high school in the week leading up to prom. The game was produced in the Expressive Intelligence Studio at UCSC, and was both nominated for national independent game awards, and instrumental in basic research into the development of "social physics" in which the narrative systems in the game dynamically reacted to player decisions and altered the game state to reflect the ever-changing relationships between the player and non-player characters.

The grant work involved a full appraisal of *Prom Week*'s development documentation, including 14 hours of interviews with the development team, and a thorough content analysis of the team's shared development servers and code repositories. The final project report, summarized in [6], outlined the issues inherent to game development documentation analysis and provided recommendations for future projects hoping to produce more "archive-ready" material collections. The analysis structure was based on previous archival work in the history of science and technology that aligned with the technical documentation present in the Prom Week project [7].

This ETC work can be seen as a significant extension to our previous work in cultural software research, as it casts a wider net toward more varied interdisciplinary media projects of, in some cases, significantly greater scale. That said, in terms of document variety and organization, as noted below in our brief content analysis, it appears that the organization of code and assets in ETC's projects is in many ways similar to *Prom Week*'s. Therefore, the previous analysis of *Prom Week* is more than capable of providing a blueprint for larger scale archival content analysis.

B. Meretzky Game Development Archive

The collections of Steve Meretzky at Stanford University Libraries represent one of the only collections of game development documentation, in the United States, to receive full archival treatment at a major institutional archive [8]. Steve Meretzky was one of the principle game designers at Infocom, a well known text adventure company active throughout the 1980s. Among its best known works are the Zork series of fantasy titles, and its collaboration with Douglas Adams on the official *Hitchhiker's Guide to the Galaxy* adventure game. The Meretzky collection includes a significant amount of physical records aligning with his game development work at Infocom, and at further entertainment companies throughout the 1990s and early 2000s. Stanford's treatment of this collection was a major impetus for the NEH work above, as there appeared to be little to no archival guidance on game development production records aside from Megan Winget's work on development documentation appraisal in the early 2010s [9]. In fact, a majority of Meretzky's digital content did not receive explicit archival organization beyond data ingest due to the nascence of digital preservation activities at that time. However, experience with the thorough archival treatment of game development records proved useful in adapting lessons learned to the follow on NEH work, and again to work with ETC's data, as some Meretzky documentation is contemporaneous with early ETC projects.

IV. Methods and Discussion

A. Data Sample Organization

The sample project set contained four projects from the Fall 2006 semester at ETC. Initially, the selection was based on the incorrect assumption that the earlier projects would contain less data and be more amenable to preliminary analysis. As noted below, one project turned out to be 3D animation-based with a file count and data footprint much larger than the three other game development projects combined. The four projects were, briefly:

1) *Skyrates*, a prototype for a sporadic game experience in which player's interact in a fictional world of Sky Pirates that intentionally limits player interaction time.
2) *CERT: Community Emergency Response Team*, a simulation game developed to train system administrators in dealing with insider attacks on organizational networks. An insider in this context is a disgruntled employee or contractor.
3) *Jukebox*, a collection of demonstrations for games that react in real-time to user chosen music.
4) *Granny*, a 5-minute 3D animated short film project.

Although ETC's website lists nine projects for Fall 2006, only four projects were identified in the folder structure under the Fall 2006 semester [10].

The larger backup server includes significantly more documentation of ETC's activities, including website backups and full code repositories. While those may be in scope for future analysis, the already copious amount of direct, project produced data lent itself well to the prospective historical, archival, and computational analysis outlined below.

B. Content Analysis of Sample Projects

An initial content analysis of the sample projects was conducted using common Unix command line utilities (`du`, `file`, `grep`, `cp`) and the UK National Archives DROID file format analysis tool. Custom Python 3.7 scripts in a Jupyter notebook were then used to format the DROID data into a small SQLite3 database. This database allowed for the DROID summary information to be easily queried. Files were copied from the backup server using Unix's standard `cp` facility, including flags (`-pRnv`) to preserve file modification dates and permissions. After copying, the directories were set to read-only status to preserve modification dates.

The three game projects (*Skyrates*, *Jukebox*, and *CERT*) made use of a variety of development frameworks. *Skyrates* and *CERT* were primarily Adobe Flash-based, however due to the nature of sporadic games *Skyrates* also had prototypical implementations for mobile text-based services, *AOL Instant Messenger*, and *Google Desktop*. This was to allow for the *Skyrates* gameplay to permeate into everyday working environments. *Jukebox* made use of Microsoft's XNA Game Studio framework, and *Granny*, the sole animation project, was rendered with Pixar's *Renderman* software. Of note is that the development frameworks used for all of the game projects are either discontinued or, in the case of *Adobe Flash*, to-be discontinued.

File and folder organization is commensurate with game project organizational structures outlined in part III above. No two projects used the exact same folder naming scheme or file naming conventions, however code and art assets were generally delineated. As far as basic statistics, the four project compromise 293 gigabytes (GB) of data, with *Granny* taking up 275gb (or 94%) of that total, followed by *Skyrates* (11.9GB / 4%), *Jukebox* (4.2GB / 1.4%), and *CERT*(1.9GB / 0.6%). Similarly, the number of total files was 136,002, with a respective breakdown of 109089 (80.2%), 18639 (13.7%), 6093 (4.4%), and 2181 (1.6%).

Due to the size of the *Granny* project, for this work-in-progress it was not possible to download its entire data set from ETC's backup server. In fact, even figuring out how to organize and transfer the data from ETC's backup server has proved to be logistically difficult, thus the use of four sample projects.[1] The rest of the statistics in this section are based on *Skyrates*, *Jukebox*, and *CERT*.

The project data contained 222 total discrete file extensions, 119 of which were not in the PRONOM file format registry (even in extension form). Overall 78.4% of the files present were identified, with 144 distinct PRONOM format signatures and 49 distinct MIME types. There were also 552 executable files, and an additional 785 Adobe Flash executables (.swf files), which is unsurprising given that it was the output format for all three projects. Based on MD5 hashes, around 10% (or 1.8GB) of the data consisted of redundant files, with a particular PNG in *CERT* having 78 differently named copies. The top 20 most prevalent file formats in number and size are shown in Table I and Table II respectively. SWD files are Adobe Flash data files used during development but not in the resulting, compiled SWFs. This is likely why PRONOM did not identify them.

TABLE I
FILE FORMATS BY NUMBER

Extension	File Format	Version	Count
png	Portable Network Graphics	1.1	5099
htm	Hypertext Markup Language		2543
gif	Graphics Interchange Format	89a	1943
swd			1525
png	Portable Network Graphics	1.0	1515
jpg	JPEG File Interchange Format	1.02	1472
jpg	JPEG File Interchange Format	1.01	1249
fla	OLE2 Compound Document Format		866
xml	Extensible Markup Language	1.0	802
rm	RealMedia		710
class	Java Compiled Object Code		595
mp3	MPEG 1/2 Audio Layer 3		570
exe	Windows Portable Executable	32 bit	485
psd	Adobe Photoshop		473
htm	Hypertext Markup Language	4.0	463
doc	Microsoft Word Document	97-2003	459
xls	Microsoft Excel 5.0/95 Workbook	5/95	437
skb	SketchUp Document		429
swf	Macromedia Flash	6	381
txt	Plain Text File		372

TABLE II
FILE FORMATS BY AGGREGATE SIZE

File Format	Version	Size(MB)
Adobe Photoshop		3599
(.swd)		2452
Windows Portable Executable	32 bit	2194
OLE2 Compound Document Format		1191
Audio/Video Interleaved Format		1024
SketchUp Document		923
MPEG 1/2 Audio Layer 3		739
Windows Cabinet File		543
Portable Network Graphics	1.1	500
Broadcast WAVE	0 PCM	401
JPEG File Interchange Format	1.02	368
Tagged Image File Format		341
Adobe Illustrator	12.0	273
JPEG File Interchange Format	1.01	268
Truevision TGA Bitmap	2.0	174
Portable Network Graphics	1.0	151
Encapsulated PostScript File Format	3	124
Acrobat PDF 1.4 - Portable Document Format	1.4	115
Apple Disk Image		95
Windows Portable Executable		86

Some interesting initial observations are:

1) The scale of the Granny animation project, with commensurate intermediate rendering stages and scene-level

[1] *Granny*'s basic statistics on data size and file number were gathered with "du" and "rsync" running through a remote connection.

draft videos, dwarfs the data footprint and file numbers of the game projects.

2) Organization within projects is very much up the individual teams, however art and code are generally distinguished in the file hierarchy.
3) There is significant and constant production of intermediate results in the form of test prototypes and test animations. Due to ETC's expressly iterative nature this is to be expected, but it is also confirmed in the file system structures themselves.
4) The second largest format by size and fourth by number of files is a development only artifact. An observation that validates further format investigations of development archives.

It is unknown, without further analysis, whether the above observations will pan out for the rest of the data set. Given the size of the ETC archive, and its 20 year time scale, it is reasonable to assume that development project management and data management processes evolved over that interval.

C. Historical Methods and Research Questions

A major challenge in the historical study of any science or technology field is the lack of access to internal project documentation that might shed light on what Science and Technology Studies (STS) scholar Henry Collins refers to as the "tacit knowledge" embedded within a practitioner's practice [11]. This knowledge does not manifest directly in an object, in our case here an entertainment product, but is essentially for the object's creation and success in its given domain. The object is therefore what's known as a "black box", a production of technology whose internal workings are hidden from view by those who interact with it. The potential for development documentation analysis is as a means for prying open the box and investigating the formative assumptions made in its creation. Documentation can provide a view into the tacit processes of production and direct critical inquiries in ways not possible through the sole analysis of an objects surface. To align with our project here, the production of games and novel entertainment systems function within a larger ecosystem tied to capital processes of "innovation", technological "progress", and the definitions of what constitutes a "game" or "interactive experience". As ETC's faculty are primarily aligned with commercial production industries, and given that a majority of ETC's student translate their experiences in the program into jobs within entertainment industries (over 90% of graduates end up there), the documentation present in the ETC archive forms a substantial base of data on which to construct inquiries about media arts productions over the last 20 years. Below we elaborate on prospective research questions, including how they could arise from the data and their implications for historical inquiry into software development process.

1) How have development processes in interactive media arts changed over time?: Diachronic analysis of a software production environment of singular provenance is rare. The organization and maintenance of ETC's data could allow for a deep historical investigation into the changes in development processes, software architectures, and project management over the last 20 years. Since the ETC is generally at the forefront of new technologies, the project data provides a contemporaneous record of innovative technologies at that time along with the development documentation to allow for their investigation.

The ETC's development processes have been the subject of numerous studies into design team organization and production processes [12], [13]. Software development methodologies could be analyzed in a similar vein, using the source material here as a guide for interviews or inquiries with ETC faculty and former students.

2) What new historical research insights can be gleaned from more complete access to project documentation?: As there is relatively little extant development documentation for technological entertainment projects, there is a potential for new, documentation-based study and historical visualization of development process. Given the "big data" represented by this ad hoc archive, implications for "big data" software histories are potentially significant.

3) How do innovation narratives in project implementation affect design choices?: ETC is explicitly an "innovator" in the space of media arts, with a significant number of project aligned with industry clients looking for pilot uses of new technologies. How does the ethos of ETC affect its development processes and the outlook of its students and faculty? Can more thorough analysis of project documentation shed light on what "innovation" means in the space of games and entertainment productions?

4) What are the cumulative effects of ETC's influence on the larger industry ecosystem?: Do ETC's projects resonate within the larger industry? Would it be possible to determine if methodologies and processes from ETC affect projects pursued by graduates and faculty after their involvement with the program? In STS there is a notion of "technoscience", which states that the fields of science and technology are related and reinforced in their growth and advancement through internal and external socio-political contexts [14]. Understanding "technoscience" requires access to the fundamental assumptions being made about how science and technology are constructed; it is the froth of uncertainty and negotiation that coalesces into scientific proofs and solidified technical objects. Critically, understanding construction of science and technology is based on examination of the intermediate and not-yet-settled "technoscience", which is in turn based on access to records and documentation of unfinished products and still-unproved claims. In this way, development documentation can shed light on a "technoentertainment", in which the advances in entertainment technology products are related to the internal design and technology decisions within a discipline. Additionally, the larger influence on entertainment, writ large, can then be interrogated through the lens of its fundamental "technoentertainment" constructions.

D. Archival Methods and Research Questions

Archival science is still struggling with the breadth and complexity of software preservation, and born-digital technical production documentation. The questions below outline how large collections of development documentation could improve the field and provide guidance for analysis and archival treatment of similar collections in the future.

1) What is the general constitution of entertainment technology development project collections?: As revealed in the content analysis in the next section, the projects in the ETC archive represent a variety of different technologies, development constraints, and outputs. The breadth of the projects, in total, may shed light on the construction of similar contemporaneous projects outside ETC, and allow for more refinement of the appraisal strategies described in our previous work. The ETC data is multiple orders of magnitude larger than any other known collection of this type, and could allow for more informed inquiries into other similar collections from academic and commercial sources. Furthermore, in conjunction the computational analysis described in the next section, the ETC data could provide significant support for improving digital preservation tools used by archivists.

2) How are current tools able to deal with the variety of digital objects associated with interdisciplinary media arts projects?: What are the limitations of current digital preservation tools and techniques in the context of analyzing and parsing interdisciplinary game and entertainment project documentation on the order of terabytes? Do our approaches scale well with larger, highly varied collections?

3) How can analysis improve common archival needs in file format identification, content discovery, reproduction and access to legacy software environments?: The variety of content in the ETC archive, including its numerous executable development frameworks, prototype system implementations, source code, assets, and clarifying documentation could allow for significant improvement in content exploration and elaboration tools for highly variegated born-digital collections. In many born-digital archival collections, the forms of documentation tends to be more singular, collections of film, photos, or digital textual documents. In ETC's case, each project is dealing with potentially new technologies and presents unique challenges and considerations for the management of heterogeneous file formats. Additionally, much of the content in development archive is itself executable, and if recovered and represented through re-compilation or emulation may shed new light on both processes for archival executable content and the implications of executable access.

Furthermore, the need for interpretation of the project files argues for the need to record and preserve the software that produced the intermediary, prototypical systems in the first place. A significant issue for the preservation of the NEH *Prom Week* data mentioned above was the difficulty in tracking down contemporaneous development environments and artistic software (*Adobe Flash Builder* and *Adobe Photoshop* primarily) capable of rendering and running project data. Alarmingly, the *Prom Week* project was still in-progress when the study began. We believe a major difficulty for the ETC data will be ascertaining and locating software to interpret "cutting edge" projects from which we are remove by a decade or greater.

E. Computational Methods and Research Questions

This last methodology section aligns our work with the computational agenda of computational archival science. The ability to parse and make sense of the data in the ETC archive will be based, in some part, on making use of computational methods borrowed from computer science and big data applications. This include model training for machine learning inquiries into project structure, file format identification, and semantic topic modeling, along with more common, aggregated statistical analysis of ETC's projects data in both synchronic and diachronic contexts.

1) Given the size of the archive, what methods are applicable and how can they support the research questions above?: A major task in our coming work is to find and apply computational means for the comparative and temporal analysis of ETC's projects. Since the total data collection encompasses a couple hundred projects totalling around 30 terabytes of data, significant effort will be needed to parse the projects' file hierarchies, enumerate and describe their files, locate resources to answer historical and archival inquiries about project process and documentation, and also just logistically manage the data set for segmented analysis.

The structure of the ETC lends itself to interrogation along numerous axes relevant to software engineering inquiries. Each project has a significant code-base, iterative development approach, and novel technical concerns, and could provide significant insight into a longitudinal study of entertainment and game system design. Methods in the static analysis of code structures, comparison of project hierarchies and organization, and other document structural metrics could allow for conclusions about iterative, creative arts development along similar lines to the team dynamic studies mentioned above.

2) Can this archive tell us about software evolution and maintenance?: Two areas of software engineering research that would immediately find use for the ETC archive materials are most likely software evolution, which tracks the changes in software development architectures over time (and instruments solutions to issues of knowledge transfer and migration), and software maintenance, a field concerned with keeping older systems running and retaining the ability for institutions and organizations to maintain aging infrastructures. The alignment of the ETC data set with a specific development methodology, time frame, and set of output goals would make a ripe target for understanding the change in software development practice, over-time, on game and entertainment projects.

3) What could training machine learning and semantic models on this data reveal?: As mentioned in the foundational CAS paper [15], there are numerous machine learning and semantic modeling methodologies that would be exciting to apply to the ETC archive data, including topic modeling, graph-based network analysis of project file hierarchies, and semantic tagging and organization of project files. Could the

ETC provide a good, base training set for file format identification in games and entertainment arts, as called for generally in [16]? Feature selection aligned with a project's graph structure, or modification times, might be able to identify specific categories of projects that would then be applicable to larger, non-ETC sets of documentation. The ETC itself is interested in means for automatic classification and semantic description of past project work as a historical resource for current student projects. Finding ways to map emerging semantic controlled vocabularies for game development records, like [17], may prove fruitful in this regard.

V. Conclusion

As elaborated above, the ETC archive represents a significant potential advancement in the interpretation and analysis of game and interdisciplinary media arts projects, both for digital humanist and archival analysis. This work-in-progress is currently limited in scope to the subset of data easily downloadable from ETC's servers. We have received a complete physical transfer of the archive's data and are current working on storage organization and backup logistics. This should allow for a much larger version of the preliminary analysis presented above that alone would improve understanding of ETC's projects, both for our own research purposes, and ETC's needs in supporting this documentation as a resource for future students.

Acknowledgment

The author would like to thank Drew Davidson, Director of the Entertainment Technology Center at Carnegie Mellon University, for permission to analyze their projects. Additional, substantial thanks go to Steve Audia, head of IT for ETC, who is the main individual responsible for the articulation, construction, and maintenance of ETC's project archive.

References

[1] D. Davidson, *Creative Chaos*. Pittsburgh, PA: ETC Press, May 2017.

[2] "UT Videogame Archive - Mission." [Online]. Available: https://www.cah.utexas.edu/projects/videogamearchive/index.php

[3] "Archival Collections," Jan. 2019. [Online]. Available: https://www.museumofplay.org/collections/archival-collections

[4] H. M. Chandler, *The Game Production Handbook*, 3rd ed. Burlington, MA: Jones & Bartlett Learning, Mar. 2013.

[5] E. Kaltman, N. Wardrip-Fruin, H. Lowood, and C. Caldwell, "A Unified Approach to Preserving Cultural Software Objects and their Development Histories," 2014. [Online]. Available: http://escholarship.org/uc/item/0wg4w6b9.pdf

[6] E. Kaltman, N. Wardrip–Fruin, H. Lowood, and C. Caldwell, "Methods and Recommendations for Archival Records of Game Development: The Case of Academic Games," *Proceedings of the 10th International Conference on the Foundations of Digital Games*, 2015.

[7] C. A. Elliott, *Understanding progress as process: documentation of the history of post-war science and technology in the United States*. Society of American Archivists, 1983.

[8] "Guide to the Steven Meretzky papers relating to computer game design and interactive fiction history, 1978-2009 M1730." [Online]. Available: https://oac.cdlib.org/findaid/ark:/13030/c8862jnz/

[9] M. A. Winget and W. W. Sampson, "Game Development Documentation and Institutional Collection Development Policy," in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, ser. JCDL '11. New York, NY, USA: ACM, 2011, pp. 29–38. [Online]. Available: http://doi.acm.org/10.1145/1998076.1998083

[10] "Past Projects | Entertainment Technology Center." [Online]. Available: https://www.etc.cmu.edu/play/past-projects/

[11] H. Collins, *Tacit and Explicit Knowledge*, reprint edition ed. Chicago; London: University Of Chicago Press, Dec. 2012.

[12] K. T. Goh, P. S. Goodman, and L. R. Weingart, "Team innovation processes: An examination of activity cycles in creative project teams," *Small Group Research*, vol. 44, no. 2, pp. 159–194, 2013.

[13] M. de Vaan, B. Vedres, and D. Stark, "Disruptive Diversity and Recurring Cohesion: Assembling Creative Teams in the Video Game Industry, 1979-2009," *SSRN Electronic Journal*, 2011. [Online]. Available: http://www.ssrn.com/abstract=1954019

[14] B. Latour, *Science in action: how to follow scientists and engineers through society*. Cambridge, Mass.: Harvard University Press, 1987.

[15] R. Marciano, V. Lemieux, M. Hedges, M. Esteva, W. Underwood, M. Kurtz, and M. Conrad, "Archival Records and Training in the Age of Big Data," in *Re-Envisioning the MLIS: Perspectives on the Future of Library and Information Science Education*, ser. Advances in Librarianship, May 2018, vol. 44B, pp. 179–199.

[16] 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, pp. 16–26, Mar. 2012. [Online]. Available: http://www.ijdc.net/index.php/ijdc/article/view/201

[17] J. H. Lee, "A Conceptual Data Model and Schema for Curating Collections of Video Game Development Artifacts," 2018. [Online]. Available: https://www.imls.gov/sites/default/files/grants/lg-86-18-0060-18/proposals/lg-86-18-0060-18-full-proposal.pdf