Evaluating a Taxonomy for   
Video Game Development Artifacts

*Taxonomies for New and Innovative Domains*

| **Marc Schmalz** | **Kylie Snyder** | **Lidia Morris** |
| --- | --- | --- |
| *University of Washington*  *USA*  *mschmalz@uw.edu*  *0000-0003-1027-696X* | *University of Washington*  *USA*  *ksnyd@uw.edu*  *0000-0002-3374-2364* | *University of Washington*  *USA*  *ljmorris@uw.edu*  *0000-0002-4702-6975* |

| **Corey Cherrington** | **Tara Disher** | **Jin Ha Lee** |
| --- | --- | --- |
| *University of Washington*  *USA*  *cherri93@uw.edu*  *0000-0001-6965-9987* | *University of Washington*  *USA*  *tdisher@uw.edu*  *0000-0002-3285-1411* | *University of Washington*  *USA*  *jinhalee@uw.edu*  *0000-0002-9007-514X* |

**Abstract – Digital game development is innovative and intersectional, producing cultural texts in an emerging field across new technology, physical, and digital media. As such, it offers fertile ground for designing and evaluating structures to help creators, information professionals, and others organize and preserve new domains, and to expand the processes of knowledge organization. Participants classified digital game development artifacts from one online and two physical archives. Data were analyzed with mixed methods, generating recommendations for improving the taxonomy and insights on evaluation framework.**

**Keywords – Metadata, Games, Taxonomy Evaluation**

**Conference Topics – Innovation**

# Introduction

Institutions such as the Strong National Museum of Play, the Stanford Libraries, and the National Media Museum in the UK now seek to catalog, classify, and preserve digital games, but are primarily focused on preservation of the final product. Less consideration has been given to the artifacts associated with their development, materials that future researchers, historians, and professionals will rely upon. Development artifacts are vital for study of the medium, helping us understand game design, intended audience, public reception, and impact on the parent organization. Today, many of these artifacts are born-digital, facing a new set of challenges for archiving. Without organized efforts to preserve such materials, they will be lost.

Researchers at the University of Washington Information School received a National Leadership Grant from the Institute of Museum and Library Services (IMLS) in 2018 to “create a conceptual data model and metadata schema for describing and representing artifacts related to the development of digital games” [1, p. 1]: the Taxonomy of Video Game Development Artfacts (TVGDA). Evaluation of the TVGDA, as a newly developed taxonomy, is ongoing. We contribute to its evaluation by having users apply it to real-life collections and evaluating the results.

There is limited prior research on evaluating methods for taxonomies in library and information science literature. Reference [2] does provide a set of qualities by which a taxonomy may be evaluated (concise, robust, comprehensive, extendible, and explanatory) and that serves as a frame for analysis.

Stated formally, our research questions are: 1) In what manner does the TVGDA exhibit the qualities of the criteria suggested in [2] for evaluating taxonomies?; 2) What suggestions can be made to further improve the TVGDA?; 3) How can this evaluation of the TVGDA inform and improve upon taxonomy evaluation processes?

Analysis of the TVGDA based on user feedback informs innovative strategies for preservation of materials about digital interactive games, as the TVGDA provides a controlled vocabulary (CV) capable of describing this unique set of information objects known as video game development (VGD) artifacts.

# Literature Review

Though suggested best practices for CVs and metadata exist, there is no common standard for taxonomy quality and few specifics for testing CV efficacy. “Most evaluation seeks to identify and improve metadata quality, but few attempt to define concretely what ‘quality’ entails” [3, p. 3]. Due to this lack of literature, many taxonomies remain unevaluated [4]. Reference [4] breaks down the evaluative practices discovered through their literature review into five categories: Logical Argument, Expert Evaluation, Action Research, Case Study, and Illustrative Scenario.

Reference [2] recommends a set of attributes by which taxonomies might be evaluated, proposing that good taxonomies are concise (limited enough in detail as to afford easy use), robust (complex enough to differentiate between objects), comprehensive (able to address any object in the domain and/or addressing all aspects of objects in the domain), extendible (able to include new dimensions), and explanatory (“provide useful explanations of the nature of the objects under study” [2, p. 342]). Specific tests for these attributes are not offered. Instead, the authors point out that most appropriate methods for evaluating a taxonomy are dependent on how the taxonomy will be used as implemented. While best practices are likely to emerge as taxonomies are improved and tested over time, there will likely remain no one-size-fits-all approach.

1. *Considerations for Classifying VGD Artifacts*

For audio and video, organizations such as the Association of Moving Image Archivists (AMIA) assist archivists in describing “Moving Image and Sound Collections” [5]. While these collections have received more attention in recent years (c.f. [6]), existing guidelines lack information on video games and their development artifacts. Reference [7] finds that the closest conceptual standard for the description of VDG artifacts is in Describing Archives: A Content Standard (DACS) [8], noting there have not been enough accessions of VGD artifacts performed to properly judge its appropriateness. This assertion necessitates content standards specific to VGD artifacts. VGD projects, while creating digital products, began before the “digital revolution… when email blasts replaced circulated paper” [9, p. 85; 10], so their legacy artifacts are both physical and born-digital. The industry has obvious archiving needs for born-digital artifacts, which is an area of classification for which DACS may not yet be well equipped [7]. There continue to be several types of artifacts involved in VGD that are in physical format, too. Thus, tools designed for video game archivists and historians need to consider born-digital, physical, and digitized artifacts. The TVGDA framework is intended to remain stable as technology evolves while providing enough context in scope notes that new forms of VGD artifacts will be classifiable.

1. *User-focused Research Principles*

Preserving video game information through metadata is a massive challenge [11]. The more we understand video games, the more we recognize the difficulties of applying current standards and rules to describe them, and even more so for VGD artifacts.

There have been initiatives to improve the organization and description of games and VGD artifacts. For example, [12] established the Video Game Metadata Schema (VGMS) as a “list of elements which form a metadata schema for describing video games.” Reference [13] utilized user interviews to “derive and discuss key design implications for video game information systems [(IS)]…” [p. 833] to improve game-related IS. These user-focused research projects form the basis of similar metadata and taxonomical structures related to video game digital assets and ephemera.

The TVGDA is one such example, specifically targeted at organizing and describing VGD artifacts. The TVGDA was created for three classes of users: industry professionals, information professionals, and game researchers. It has a single dimension, used to describe a VGD artifact’s type. Taking its warrant from industry use, the TVGDA “is organized into three broad sections including (a) Development (with seven subsections), (b) Organization-Related Materials, and (c) Marketing (with four subsections), representing different aspects and timelines of game development” [7, p. 548]. It includes 123 industry terms with scope notes and additional lead-in terms.

# Methods

​​Per [8], TVGDA evaluation began by applying the taxonomy to 1,000 VGD objects, supplemented by two expert evaluation interviews. Data gathering took place as a graduate-level class cataloging assignment based on an online archive of VGD artifacts and a pair of individual tests on institutional collections. These methods allowed us to test consistency of use as well as applicability across multiple collections. The expert evaluation interviews were conducted as follow-up interviews on the latter two collections.

# Findings and Discussion

We used the criteria in [2] to frame our evaluation: We consider whether the TVGDA is comprehensive, concise, robust, explanatory, and extendible.

1. *Comprehensiveness*

Comprehensive is the quality of covering all objects in the intended domain, or of including enough dimensions to describe the domain [4].

**Coverage.** At 123 terms, the TVGDA is an extensive representation of VGD artifacts. Analysis shows far more difficulty deciding between terms than finding an applicable term. That said, results did include indications of possible missing terms.

There were more than 20 comments from graduate catalogers requesting or suggesting new terms, often narrower terms for artifacts with niche purposes. These suggestions usually came from working with difficult-to-describe materials, including game control or navigation graphics, tables of contents, barcodes, and physical comic books. Still, these suggested terms may be requesting a level of specificity that may not be necessary in the TVGDA.

**Dimensionality.** Comprehensive taxonomies should contain enough dimensions to describe the domain. Specifically, [2] says: “a useful taxonomy includes all dimensions of objects of interest” [p. 341]. This is differentiated from robustness, defined as, “enough dimensions and characteristics to clearly differentiate the objects of interest” [p. 341]. Comprehension and robustness must be balanced in each taxonomy. The TVGDA is a unidimensional expression of an artifact’s type in a highly specific domain, but institutional implementations would certainly include other dimensions (also present in the student cataloging exercise). We continue the analysis of dimensionality below, under Robustness.

1. *Conciseness*

Concision is the quality of parsimony and limited complexity and is at tension with the quality comprehensiveness [2]. Graduate catalogers were not asked to review this quality of the TVGDA but the research team notes that the TVGDA prefers comprehensiveness to concision.

The sheer number of terms may also be a factor in relatively low intercoder agreement (see Explanatory Power, below) as catalogers fall back on familiar terms where less-used terms may be more appropriate. The TVGDA’s caretakers should consider user studies to reduce the number of broad terms, easing the conceptual load required for high-level classification of artifacts while allowing interested parties to use narrower terms for detail.

1. *Robustness*

Robust is the quality of allowing catalogers to differentiate between objects: “enough dimensions and characteristics to clearly differentiate the objects of interest” [2, p. 341].

Test catalogers had problems choosing terms in this unidimensional taxonomy. For example, participants vacillated between *screenshot* (“Image captured from a game during play” [14, p. 6] and *art asset* (“Any artwork used in a released version of the video game, such as 3D models or 2D artwork” [14 p. 5]). The intended purposes of an image may be required to determine whether any given image is a *screenshot* or one of the other graphical artifact types.

The all-digital online archive highlights another issue with images: Catalogers sometimes classified the artifact as an image and sometimes as the object depicted. Cataloging standards help information professionals understand how to classify these objects, but the TVGDA is intended to be used by creators as well.

To make the TVGDA more robust, the research team suggests that its caretakers consider separate terms for nature and function, add additional guidance on which aspect should take priority, consider guidance on the use of multiple terms for single artifacts, and clarify the function of digital representations of physical artifacts for creators.

1. *Explanatory Power*

Explanatory is the quality of adequately describing the domain, providing “useful explanations of the nature of the objects under study” [4, p. 342]. We compared the behavior of multiple catalogers to see if they shared a common understanding of the items described. Quantitative analysis provides insights.

Since the artifacts were distributed to multiple groups who further distributed the work to members, assumptions for use of Cohen’s kappa (sets being compared being completed by one and only one coder [15]) were not met and could not be used. Additionally, many students selected no terms or multiple terms, so these pairings were omitted. Analysis used simple statistics regarding matches between valid pairs, but at multiple levels.

Many catalogers labeled items as “ambiguous” and required further analysis. Participants often noted that a lack of context meant they could not accurately assign any terms. Instead, they relied on several other indicators to attempt a classification, including inferences from file names and online searches for authoritative information. While failure to agree on understood objects shows room for improvement in the TVGDA, its caretakers have little ability to control the ambiguity of the nature of a given item, or its lack of context in an archive.

1. *Extendibility*

Extendable is the capacity for a taxonomy to be revised. The TVGDA is a single-dimension taxonomy, and there are no barriers to extending it in terms of adding another identified dimension or adding sections or terms to the existing dimension. In this regard, the TVGDA is extendable.

# The Evaluation Framework and Broader Implications

Comprehensiveness represents two qualities— coverage and dimensionality—which were treated separately. We found no benefit to considering them as a single quality. Coverage was the most intuitive sub-quality for the research team to grasp and evaluate: Have participants identified artifacts which cannot be satisfactorily classified with the current taxonomy? Dimensionality is difficult to address in a unidimensional taxonomy without specific participant feedback, and we had none.

Conciseness is in tension with comprehensiveness and is tied to the cognitive load required to apply the taxonomy. Aside from the two interviews, we did not inquire directly about cognitive load, though we believe quantitative analysis of cataloger agreement offers us some insight in this work. We believe guidance for how and when to use broader or narrower terms should help situationally balance conciseness with comprehensiveness.

Robustness seems tied to multi-term classification with this unidimensional taxonomy. The volume of multi-term suggestions and lack of coder agreement seem like inverse measures for robustness, indicating room for improvement.

Explanatory power was evaluated here by quantitatively evaluating cataloger agreement, as a sufficiently explanatory taxonomy will provide a common understanding of the domain and make classification easier. This ties explanatory power to robustness: Lacking full explanatory power in taxonomy, our users often found that additional term selection (tied to robustness) helped them explain an artifact. This is complicated by the presence of ambiguous artifacts: those whose nature is explained through the taxonomy but not understood by evaluators due, perhaps, to lack of experience in the domain or with the specific collection.

Many concerns found in the research seemed to have multiple modes of evaluation. Troubles identified with *screenshot* could situationally apply to robustness or comprehensiveness, for example. Still, the team found these qualities to be useful evaluative concepts and intend to use them in the future to help researchers formulate better methods for their evaluations for any given specific taxonomy.

# Conclusion

Our approach to assessing this taxonomy combines the meaningful elements of established heuristics and user testing to create situated vocabularies and taxonomies to establish best practices for defining the breadth and depth of relevant artifacts across evolving domains. The TVGDA offered the opportunity to contribute through evaluation of both a new user-centered taxonomy and a framework of taxonomy qualities, making recommendations for improvement of both. Focus on the other two user groups (creators and researchers) represent an additional opportunity for expanding tests of the TVGDA. VGD is constantly evolving and intersectional, producing cultural texts by developing new technology and spanning physical and digital media in a relatively unexplored domain that has become academically legitimized only relatively recently. As such, it offers fertile ground for designing and evaluating structures to help creators, information professionals, and other users organize and share the domain, and to expand our knowledge of knowledge organization as well.

# Acknowledgment

This project was made possible in part by the Institute of Museum and Library Services.

# REFERENCES

1. LG-86-18-0060-18, http://www.imls.gov/grants/awarded/lg-86-18-0060-18-0, last accessed 2021/09/13.
2. Nickerson, R.C., Varshney, U., Muntermann, J.: A method for taxonomy development and its application in information systems. European Journal of Information Systems. 22, 336–359 (2013).
3. Lee, J.H., Clarke, R.I., Perti, A.: Empirical evaluation of metadata for video games and interactive media. Journal of the Association for Information Science and Technology. 66, 2609–2625 (2015).
4. Szopinski, D., Schoormann, T., Kundisch, D.: Because Your Taxonomy Is Worth It: Towards a Framework for Taxonomy Evaluation. Research Papers. (2019).
5. Cocciolo, A.: Moving image and sound collections for archivists. (2018).
6. AIMS Work Group: AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship, https://dcs.library. virginia.edu/files/2013/02/AIMS\_final\_text.pdf.
7. McDonald, C., Schmalz, M., Monheim, A., Keating, S., Lewin, K., Cifaldi, F., Lee, J.H.: Describing, Organizing, and Maintaining Video Game Development Artifacts. Journal of the Association for Information Science and Technology. 1–14 (2020).
8. Society of American Archivists: Describing Archives: A Content Standard (DACS), https://www2.archivists.org/groups/techn ical-subcommittee-on-describing-archives-a-content-standard-dacs/describing-archives-a-content-standard-dacs-second-.
9. Williams, J.A., Berilla, E.M.: Minutes, Migration, and Migraines: Establishing a Digital Archives at a Small Institution. The American Archivist. 78, 84–95 (2015).
10. Wolf, M.J.P. ed: *The video game explosion: A history from PONG to Playstation and beyond*. Greenwood Press, Westport, Conn (2008).
11. Winget, M.A.: Videogame preservation and massively multiplayer online role-playing games: A review of the literature. Journal of the American Society for Information Science and Technology. 62, 1869–1883 (2011).
12. Lee, J.H., Perti, A., Clarke, R.I., Windleharth, T.W., Schmalz, M.: UW/SIMM Video Game Metadata Schema Version 4.0. (2017).
13. Lee, J.H., Clarke, R.I., Rossi, S.: A qualitative investigation of users’ discovery, access, and organization of video games as information objects. Journal of Information Science. 42, 833–850 (2016).
14. Lee, J.H., McDonald, C., Schmalz, M., Windleharth, T., Keating, S., Monheim, A., Cifaldi, F., Lewin, K.: Taxonomy of Video Game Development Artifacts. (2020).
15. Hoyt, W.T.: Interrater Reliability and Agreement. In: Hancock, G.R. and Mueller, R.O. (eds.) The reviewer’s guide to quantitative methods in the social sciences. Routledge, New York (2010).