Duct-Taping Databases, or How to Use Fragmentary Online Data for Researching “Japanese” Videogames

Peter Mühleder, Martin Roth, Tracy Arndt, Florian Rämisch

Abstract
The internet provides exciting opportunities for scholars of Japanese videogames. Fans from all over the world passionately collect, order, and share information about their hobby online. In this paper, we will show how we can exploit these potentials for research on Japanese videogames. We call this approach “duct-taping” databases: the integration of information from various heterogeneous, fragmentary online resources with the goal of creating a robust research dataset. We will discuss the potentials and challenges of this approach and show how it allows us to better understand the historical development of Japan’s videogame production.

Keywords: video games, game studies, data integration, data matching, digital methods

1. Introduction
The term “Japanese videogames” is as ambiguous as it is popular. Putting problematic essentializations of “THE Japanese game” aside, games may be labeled “Japanese” for various reasons: because they have been produced in Japan; because they are sold or popular in Japan; or because their content or aesthetics are perceived as Japanese or ascribed to Japan in some way (see Goto-Jones 2015). Each of these categories is vague in one way or another. Regarding the framing of Japanese games as games produced in Japan, Mia Consalvo remarks:

1 Sächsische Akademie der Wissenschaften zu Leipzig, Germany
2 Ritsumeikan University Graduate School of Core Ethics and Frontier Sciences, Japan
3 University Library, Leipzig University, Germany
4 University Library, Leipzig University, Germany
It no longer makes sense (if it ever did) to talk about “the Japanese game industry” or even “the game industry” as if it were a monolithic entity. Platforms have expanded, the game-playing demographic is wider than ever before…. Game developers can work for large multinational corporations like Square Enix and Nintendo, or they can work in an apartment and sell their game at a site like the Tokyo Comic Market or online. Either version now also has the chance to be localized and sold in multiple markets outside the game’s country of origin. Growing industry segmentation makes sense and also signals a maturation of the market. (Consalvo 2016, 2)

Consalvo has also perceived a trend toward globalizing within Japan’s videogame industry. Analyzing the strategies of several influential Japanese companies, she concludes that “[e]ven if revenues are comparatively small, the global operations of Japanese game companies have been key to the development of the industry generally, as Japanese games and firms have contributed much in the way of cultural (if not economic) influence on the global games industry” (Consalvo 2009, 139). Mark Wolf notes that “just as many national videogame histories are finally being written, the growing shift toward transnational game development is eroding and reconfiguring the very concept of a national industry” (Wolf 2015, 12).

However, game industry researchers like Aphra Kerr question whether the videogame industry is actually globalizing. Summarizing her findings, Kerr states that “[c]ontent production is predominantly located in industrialised and developed countries in the global north, although there are interesting pockets of development elsewhere” (Kerr 2017, 104). She adds that “[l]ocal, regional and national contexts still exert their influence on the production and consumption of digital games, but it is a complex tapestry to unravel” (138).

This is even more the case given that reliable data about these developments is largely lacking or hidden behind company doors (Kerr 2017, 31). Our aim is to find a workaround for this problem, by drawing on several online resources to generate a dataset that allows us to better understand the historical development of Japan’s videogame production and its relation to the local and the global. Generally speaking, the internet provides exciting opportunities for scholars of popular culture and media. Online industry marketing and the proliferation of online encyclopedias and databases created by fans provide rich possibilities for accessing detailed information on a wide range of media cultures. This is also true for the field of Japan’s popular culture, which has been rigorously documented by enthusiast
communities around the world in an array of ways. Such community projects can be seen as a *vernacular practice*—that is, “everyday forms of ‘cultural participation and self-representation’ … outside the institutional contexts of galleries or museums” (Nicoll 2017, 180). These communities’ expertise, technical skills, and combined efforts have produced detailed information about anything from individual works and characters to companies and people involved in the production of popular culture. As we will show, combining such data sources provides us with a reliable—and, indeed, an unprecedented—basis for making claims about this development.

While the methods and techniques presented in this paper refer to our specific use case, we believe they can be applied to different domains, as online resources—professional as well as community-driven—grow in number. But in order to make the most of online data, we must reflect on our research methods and workflows, as they inform “…the selection, preparation, and cleaning of data, especially when data are biased or are missing” (Barocas et al. 2017, 71). This process requires making choices that will influence the analysis and, therefore, should not be taken lightly. As Solon Barocas et al. argue:

... in most cases, there are no “right” choices or even a “right” way to make these choices. Instead, researchers and practitioners almost always find themselves working through a number of difficult trade-offs between competing goals and values, trying to strike an appropriate balance for the task at hand. (Barocas et al. 2017, 71)

This paper discusses the challenges and opportunities of the approach we call “duct-taping” databases—locating relevant information in heterogeneous online sources and combining them into a research dataset. This process involves three steps, which will be elaborated in the following sections: The first section deals with identifying and comparing suitable datasets, integrating information, and dealing with dataset bias by developing sampling techniques. In section 2, we discuss the two online data sources that provide us with the crucial information necessary for our use case and discuss the conceptual problems that arise when dealing with heterogeneous data sources. Section 3 discusses the process of integrating the two datasets, and section 4 shows how we deal with data bias in our sources and how we are able to maintain empirical representability in our final research dataset. The conclusion offers an outlook on how this research dataset can be used to generate new insights on the spatial dispersion of videogame production. Finally, we reflect on the challenges and
opportunities of this research process and how the decisions in the research process informed these results.

2. Fragmentary Online Data Sources

As stated above, there are numerous online resources for videogames available, ranging from official institutions—for example, age rating agencies like the Entertainment Software Rating Board (ESRB) or the Pan-European Game Information organization (PEGI)—to community-driven projects such as wikis or fan databases like MobyGames. Given our focus on the production processes of console videogames released in Japan, we are interested in the following information:

1. All console games released in Japan.
2. The companies involved in the production of these games.

This focus on *console videogames released in Japan* is the first concession to the available data, as the most complete online resource on videogames, the Media Arts Database,\(^5\) focuses on games released in Japan and claims a high degree of completeness only for the console games sector. The database was created by the Japanese Agency for Cultural Affairs in 2010 with the intention to survey “the overall picture of media arts that have been created in Japan.”\(^6\) As such, it contains information about videogames, manga, anime, and media art. It is administered by a group of researchers from Ritsumeikan University and Tsukuba University. According to this research group, the database provides almost 100 percent coverage of all console game releases published in Japan since the 1980s.\(^7\) It is the only

\(^5\) Accessed July 27, 2020, [https://mediaarts-db.bunka.go.jp/](https://mediaarts-db.bunka.go.jp/).

\(^6\) Agency for Cultural Affairs, “About Media Arts Database (Beta Version),” accessed July 27, 2020, [https://mediaarts-db.bunka.go.jp/about](https://mediaarts-db.bunka.go.jp/about).

\(^7\) A list of all game consoles available in the Media Arts Database can be found in the DIGGR GitHub repository, accessed July 29, 2020, [https://github.com/diggr/platform_mapping/blob/master/dist/mediaartdb.csv](https://github.com/diggr/platform_mapping/blob/master/dist/mediaartdb.csv).
comprehensive resource that includes bibliographic records of videogames in Japan. As such, it can serve as a virtual reference to the basic population of our study.8

While the claim of completeness is advantageous for empirically grounded research, the available information regarding the games’ production processes (companies, staff, etc.) and other metadata is limited. We therefore needed to identify an additional data source that provides this information. The most detailed listings of company information can be found in the fan-based database MobyGames9 (see table 1). MobyGames was founded in 1999 by three high-school friends with the aim of sharing information about videogames online. The site grew into one of the biggest collections of videogame data and is supported by an active community (which gathers information and/or financial support via the content subscription platform Patreon). Unlike the Media Arts Database, MobyGames does not provide information on all videogames and there is no clear indication about which games are missing. Furthermore, among the companies listed by MobyGames are those that were only working on a specific local release of the game (e.g., companies doing translation work or local advertisements). Bearing these specifics in mind, MobyGames provides detailed, well-structured, easy-to-access information10 about the companies involved in the production of videogames.

| Yakuza 0 (PlayStation 4) |
|--------------------------|
| Media Arts Database      | MobyGames         | IGDB.com11       |
| Published by: 株式会社セガ | Published by: SEGA Corporation | Published by: Sega |

8 Because of a lack of other similarly rigorous efforts, this definition of the basic population is the only one possible given the available resources. This definition excludes PC and arcade games, as well as the indie market.

9 MobyGames home page, accessed July 29, 2020, http://www.mobygames.com/.

10 MobyGames was kind enough to provide us with a “pro account” for its API, allowing us to access more information than we would otherwise have been able to retrieve via their public API (accessed July 27, 2020, http://www.mobygames.com/info/api).

11 IGDB is the abbreviation for the Internet Games Database, a website featuring various data on games operated by Twitch.tv, which is owned by Amazon.
Combined, these two resources can provide us with the data we need for our research case: the console games released in Japan can be found in the Media Arts Database, and more detailed information about the production process is available from MobyGames (see table 2). In this case, the relationship between the two resources is clear: The Media Arts Database provides the basic population and MobyGames additional information. But this relationship is case-specific; in other cases, different combinations of resources must be considered.

| Data Source       | Game Entries | Language | Scope     | Japanese Release Date(s) | Credits | Companies | Alternative Titles |
|-------------------|--------------|----------|-----------|--------------------------|---------|-----------|-------------------|
| Media Arts Database | 38,068       | Jp       | Japan     |                               |         |           |                   |
| MobyGames         | 62,131       | En       | Worldwide |                               |         |           |                   |

Table 2. Available information from two different data sources

Data retrieved from these online sources, moreover, have a somewhat problematic ontology: data are not a neutral resource, they are “collected,” “entered,” “compiled,” “stored,”
‘processed,’ ‘mined,’ and ‘interpreted’” (Gitelman and Jackson 2013, 3). As we attempt to combine information from various online sources, we must be careful not to assume that their entries refer to the same objects. As our research shows, the two datasets differ not only in their data structure, but also on a conceptual level. In other words, while both datasets contain information about videogames, they have different views about what a game is.

The MobyGames instructions for new entries state: “Basically we’re documenting how you can obtain a game, i.e. the game was released as a vanilla version and a collector’s edition, and later in a compilation. Each of these should have its own entry...”12 Despite that, different versions of a game (e.g. releases on different platforms) are often subsumed under the same entry. The Media Arts Database, on the other hand, uses the term “game package” and defines each entry as a release: each version of a game (e.g. different console ports, digital download versions, special editions, budget re-releases) received its own entry with a unique identifier. Table 3 provides an elementary example of how the different views and definitions are expressed in data. It shows clearly that there is no one-to-one link between the Media Arts Database and MobyGames entries: MobyGames summarizes different platform editions into one entry, whereas the Media Arts Database creates single entries for each platform and even for each media type.

| Media Arts Database ID13 | Platform | MobyGames ID/Slug | Platform |
|-------------------------|----------|------------------|----------|
| 039213402217            | プレイステーション3 ディスク PlayStation 3 Disc | https://www.mobygames.com/game/yakuza-0 | PlayStation PlayStation 4 Windows |
| 039213402218            | プレイステーション3 ダウンロードコンテンツ PlayStation 3 Download | | |
| 392145100247            | プレイステーション4 ディスク PlayStation 4 | | |

12 The MobyGames Standards and Practices,” section 1.2.3, “New Entry.” Updated August 24, 2019. https://www.mobygames.com/info/standards#New_Entry.

13 This research uses the Media Arts Database identifiers which were in use until 2019.
Table 3. Example of equivalent entries for the videogame Yakuza 0 in MobyGames and the Media Arts Database

| Disc                          | PlayStation 4 Download |
|-------------------------------|-------------------------|
| 392145100248                 |                         |

Although we can create simple links between the entries of different databases, it is crucial to understand the semantic differences and nuances of the matched entities. A failure to do so weakens the expressiveness of the results as it creates a connection between corresponding entities on different conceptual levels, making the evaluation and interpretation of resulting data more difficult.

In order to avoid confusion on a theoretical level, we discussed, defined, and identified different kinds of conceptual entities for videogame objects. Subsequently, we adapted the videogame entities Game, Edition, and Local Release proposed by Jacob Jett et al. (2016) and Jin Ha Lee et al. (2014). By applying the proposed videogame entities to our data sources, we can see that, on one side, we have Edition entries in MobyGames and, on the other side, Local Release entries in the Media Arts Database. This is a crucial distinction because it directly impacts the analysis and interpretation of the data.

Overall, data about videogames is distributed over several resources. These resources differ in perspective and aim, which can result in heterogeneous syntax, schemas, or semantics. The use of such data therefore requires an understanding of data in various contexts. In conclusion, Espen Aarseth’s observation that “[w]hen two or more game researchers are using the word ‘game’, they may or may not be speaking about the same thing” (2011, 50), also holds true for the people in charge of online videogame databases.

---

14 “With regards to computer games we aim to distinguish the composite entertainment product generally referred to as game (game as object) from the socially negotiated activity such objects can support (game as process)” (Aarseth and Gordon 2015, 2).
3. Data Integration

Having identified suitable data sources, and having understood their schematic and semantic specifics, the next important step was the actual integration of these datasets. Data integration is, as Miguel Macias-Garcia, Víctor Sosa-Sosa, and Ivan Lopez-Arevalo (2009, 1) argue, a “slow, error prone and inefficient” process. A crucial aspect in this process is the identification of corresponding entries. This linking (also called matching or merging) of increasingly complex and heterogeneous datasets has become a pivotal issue in digital humanities and social research (Harron, Goldstein, and Dibben 2016a, 7). Traditionally, scientific fields such as statistics, database research, or artificial intelligence deal with this problem in different ways (Cohen, Ravikumar, and Fienberg 2003). Nowadays, terms such as record linkage, object identification, entity disambiguation, deduplication or merge/purge, and entity resolution have become commonplace in fields as diverse as computer science, marketing, libraries, and insurance (Winkler 2016, 9).

Generally, there are two methods for entry matching: deterministic linkage and probabilistic linkage. In deterministic linkage, corresponding entries are identified using strict rules, often involving the matching of unique identifiers. Consequently, deterministic matching results in a low rate of mismatches. On the other hand, there is also the potential for a high rate of missed matches in cases where unique identifiers are not available for all entries (Harron, Goldstein, and Dibben 2016a, 2–3). Probabilistic linkage, by contrast, does not require exactly matching fields, but rather calculates the probability of two records matching. Based on this probability, the comparison pairs are then categorized into true matches, non-matches, and possible matches (see Harron, Goldstein, and Dibben 2016b; Christen 2012). Probability-based matching has become “perhaps the most widely recommended approach for satisfactory linkage of data with imperfect or incomplete identifiers” (Harron, Goldstein, and Dibben 2016b, 228), but recently, new approaches—commercial frameworks (see Köpcke and Rahm 2010; Christen 2012) and experimental techniques using machine learning and/or graph-based algorithms (see, e.g., Liu et al. 2013; Kalashnikov and Mehrotra 2006)—are also being explored.

While in some cases, the linking of datasets might only be a simple comparison operation of a shared universal identifier, the matching of heterogeneous online data is often more complicated as, in many cases, they do not provide such identifiers. In the domain of
our research, there is no single universal identifier for videogames,\textsuperscript{15} which prevented us from using a simple deterministic approach. Consequently, we devised a probability-based, multistep approach that takes the specifics of our datasets into account.

The first step in developing this solution was to identify suitable data fields for comparison. We used the categorization schema put forward by Yu Jiang et al. (2014) to categorize promising fields in three categories: \textit{decisive} (e.g., unique IDs), \textit{reliable/partially decisive} (e.g., year, journal name), and \textit{useful, but not reliable} (e.g., author list), based on applicability, selectivity, and accuracy.\textsuperscript{16}

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|}
\hline
Category & Data field & Notes \\
\hline
Reliable & Platform & Medium selectivity, high applicability, high accuracy \\
\hline
Reliable & Release date & Medium selectivity, high applicability, medium-high accuracy \\
\hline
Useful & Title/Alternative titles & High selectivity, high applicability, medium accuracy \\
\hline
\end{tabular}
\caption{Evaluation of data fields used for matching (adapted from Jiang et al. 2014, 3)}
\end{table}

In our videogame datasets, there are only a limited number of data fields that are useful for matching. The main problem was the fact that our datasets are in different languages. While translations exist for the game titles in both datasets, they are not reliable.

\textsuperscript{15} The Wikidata videogames task force (accessed July 29, 2020, https://www.wikidata.org/wiki/Wikidata:WikiProject_Video_games) is currently attempting to incorporate various videogame identifiers in their dataset, but has not yet found a way to resolve the semantic differences between the datasets discussed in section 2.

\textsuperscript{16} “Applicability is the number of non-null values/number of total records. If a field has very few null/empty values, it has high applicability. Selectivity: the average selectivity of a field is 1 - (1/number unique field values). If a field value of the field is shared by only very few records, the field have [sic] high selectivity. Especially, if we say a field has single selectivity or [is] decisive, it means that any non-null value of this field is unique among records. Accuracy is the average probability of correctness of any value in the field. If a field has low accuracy, it is not a good idea to use it as a duplication indicator.” (Jiang et al. 2014, 3)
For other fields, such as developer, distributor, or genre, no such translations exist. As these fields also have a high selectivity, the amount of translation work necessary in order to be able to match these fields would be significant. This highlights an interesting challenge for matching datasets in different languages: fields with high selectivity, rather than being desirable (since they can act as pseudo-identifier), can become problematic.

In our approach, we opted for two fields with high applicability and medium selectivity as deterministic conditions in order to limit the possible candidates for matching: release date and platform. The problem of translation does not exist for the release date, and the number of different gaming platforms is small enough for manual translation. In a second step, we used the available game titles in the datasets as the basis for an “approximate string comparison” (Winkler 2016, 20).

Such a string comparison commonly employs algorithms, such as Levenshtein distance, in order to calculate the similarity of two strings. In our case, this was not viable, as game titles are complex constructs, featuring a wide range of different content, spellings, and abbreviations across different areas. Table 5 shows different variations among videogame titles. As the examples indicate, a string comparison of game titles must take into account the different romanization approaches to the Japanese language (e.g. ū / u / uu) as well as the mixed use of roman and arabic numerals.

| Media Arts Database | MobyGames |
|---------------------|-----------|
| 龍が如く 0 誓いの場所 | Yakuza 0 |
| リュウガ ゴトク 0 チカイ ノ バショ | 龍が如く 0 誓いの場所 |
| Ryu ga gotoku 0 chikai no basho | Ryū ga Gotoku 0: Chikai no Basho |
| CRISIS CORE – FINAL FANTASY VII – | Crisis Core: Final Fantasy VII |
| CRISIS CORE FINAL FANTASY 7 | クライシス コア -ファイナルファンタジーVII |
| Kuraishisu koa fainaru fantaji 7 |

Overall, there are 34 platforms present in the Media Arts DB dataset. For a list of platforms in different sources for videogame data, as well as a standardized mapping, see “DIGGR Platform Mapping,” DIGGR GitHub repository, accessed July 29, 2020, https://diggr.github.io/platform_mapping/.

A mix of roman and arabic numerals in two titles, for example, would in most cases result in a greater Levenshtein distance than two titles with different arabic numerals.
As a result, we developed a game title comparison algorithm that takes several characteristics of videogame titles into account. In addition to standard similarity measures, we extracted the following information from the title strings:
- Numbers (arabic and roman numerals and their position in the title)
- Markers indicating a subtitle
These additional data were weighted and added to the overall similarity of two game titles. The results were encouraging: About 40 percent of all entries in the Media Arts Database could be matched with a corresponding entry in MobyGames (as a possible or correct match). Compared to the number of false and correct matches, the number of possible matches turned out to be rather limited (about 4 percent).

4. The Research Dataset

By matching our two datasets, we were able to obtain additional data about game production for some of the entries in the Media Arts Database, but not all of them. This limits the number of entries in the Media Arts Database that we can use as the basis for our further research. Therefore, we must deal with possible biases in the remaining data. This can be achieved by drawing a sample, but it should be noted that this does not compensate for biases of human origin already existing in the data.

This bias can be tied to the data digitization process, which was necessary for us to be able to work with some of the data. This might sound counterintuitive, as videogames are

---

19 Published on GitHub: “Game Title Utils,” DIGGR (Databased Infrastructure for Global Games Culture Research) GitHub repository, accessed July 29, 2020, https://github.com/diggr/game_title_utils.
20 The resulting links are published on Zenodo (Mühleder, Rämisch, and Hoffmann 2019).
born digital (Manovich 2017), yet their metadata and release information are born analog (Manovich 2017), thus making this information available, but not machine readable.

These metadata—or data apparently born analog—require human interaction if they are to be transformed into a machine-readable format. Information garnered from the games—for example, which company produced a game—is often only available in-game via end credits and the digitization of this information requires a lot of effort. As this exercise can be time-consuming, owing to complexity or sheer volume, a selection process is required. Decisions are made about which games will be enriched with further, detailed information and which are to be omitted. According to Lev Manovich, digitization may amplify pre-existing biases and filters of modern cultural canons (2017).

In sum, making metadata is a curatorial process. With regard to fan-centered online communities, according to Benjamin Nicoll, this follows a different logic than “official” archives: It favors “‘expression over structure’, authenticity over impersonal content management...” (Nicoll 2017, 182) and focuses not only on “appreciating the historical content,” but also on “collaboration and community participation, where the cultural meaning ... is elaborated collectively.” In other words, “[p]layful participation ... is a large part of what makes the curatorial experience rewarding” (188).

For example, in our context, popular games may have a higher chance of being supplied with credit data than unpopular games. “Good” games (but not necessarily economically successful ones) with a supportive and strong fan base might have a higher chance of motivating enthusiastic fans to supply community databases like MobyGames with information about releases, cover images, credits, and artworks. The perception filter (or bias) from the analog world is transferred to the digital world, as “bad” (or less popular) games might not be equipped with sufficient information. Moreover, special interest games that address a very small community may be either overrepresented, as the fan base in some cases might become enthusiastic and willing to supply information to community databases, or underrepresented, because the community is, on the whole, smaller, thus there is a smaller chance statistically that a record will have been created for a given game. A larger fan base could increase the chances of a game having sufficient records. Or, in the case of indie games, an active and enthusiastic community could influence the availability of data, even though their popularity (in terms of sales) is far less pronounced. Games like Policenauts and Mother
3, for example, were never released outside Japan, but have entries in English-speaking community databases.21

The curators of MobyGames are diverse (originating from around the world, of different genders, and with various backgrounds) and their personal preferences tend to level out, making the overall selection appear to show a distribution similar to the original distribution of games in the Media Arts Database in terms of platform and release year (see fig. 1).

21Mother 3 entry in MobyGames, accessed July 30, 2020, http://www.mobygames.com/game/mother-3; Policenauts entry in MobyGames, accessed July 30, 2020, http://www.mobygames.com/game/policenauts.
Duct-Taping Databases, or How to Use Fragmentary Online Data for Researching "Japanese" Videogames
Yet, we identified some differences (see fig. 1: 1a–1b, 2a–2b, 3a–3b) where the selection made by the curators differs visibly from the original distribution of games. Because investigating these biases further was beyond the scope of this paper, we assumed that every game, in principle, had the same chance of being supplied with the information required to answer our research question.

In order to minimize the bias present in our data, we applied a stratified sampling technique to draw a sample. To generate strata to sample from, we used properties that also functioned as blocking keys during the data-matching process (release year and platform), as they were available for all entries. From this, 264 groups or “bins” (all combinations of the 34 years and 33 platforms with at least one release) were generated. The share of each group from the basic population was used to obtain a weighted probability for each group. In the final step, we categorized all matched entities (i.e., Media Arts Database releases with a match to a MobyGames edition) into the aforementioned groups, allowing us to draw a sample with the same distribution in terms of platform and release year.

Besides providing us with a reliable sample, which became our final research dataset, this approach also had the advantage of providing additional insights into the datasets. For example, visualizing the dataset as a 2D matrix of all release years and platforms, we can immediately see the lifespan and popularity of videogame platforms (see fig. 1).

But even with our careful preparation process, some sampling errors occurred. Some bins are significantly underrepresented in the sampled population, meaning that there are far fewer games in a particular bin than the weighted probability from the basic population suggests. In 20 out of 500 cases, a platform–year combination was sampled, but the corresponding bin in the sampled population was empty.

An additional problem relates to the aforementioned semantic differences between our datasets. The Media Arts Database contains all releases of a game, physical as well as digital. This distinction is rare in the MobyGames dataset, often featuring only a single

---

22 To investigate these biases, a study of common properties among those games without sufficient data could reveal commonalities between the games in question, which could be balanced by manually supplying or researching information.
release in Japan. As a result, a game that has multiple releases in the Media Arts dataset has a greater chance of being selected multiple times for the research dataset.\textsuperscript{23} With only 40 percent of all Media Arts Database releases matching with MobyGames entries, this effect is increased for the weighted sample, as each bin only holds a limited number of releases.

Despite these problems, the weighted sample provides an effective way of approximating representation of the basic population (all game releases in Japan), but not a flawless one. In future research, a more refined matching, perhaps including additional manual review, may result in a more granular disambiguation of releases, which would reduce this bias significantly.

5. Conclusion

In this paper, we have discussed our approach of “duct taping” databases: the integration of information from various heterogeneous online resources with the goal of creating a robust research dataset. This allows us to utilize the extraordinary work of videogame fans all over the world, who passionately collect, order, and share information about their hobby online. Given the lack of reliable data in the secretive videogame business, linking various databases offers a new, empirically grounded perspective on videogame culture.

This approach provided substantial research data relevant to our research interests in the development of Japanese videogame production, although it required us to shift our focus to Japan as a release region or market. Our research dataset allows us to identify a large percentage of the companies involved in the production of our sample of 480 videogames released in Japan. Moreover, we are able to locate these companies and their headquarters on a world map, thus visualizing the spatial distribution of videogame production for the Japanese market (see figures 2 and 3).

\textsuperscript{23} E.g., three out of six Xbox One games in the sample are releases of \textit{Forza Motorsport 6}; three PS3 games are releases of \textit{Lollipop Chainsaw}. 

\textit{Journal of the Japanese Association for Digital Humanities}, vol. 5, No. 1, p. 77
As the two figures show, the companies involved in producing releases for the Japanese market and their location change significantly over time. Early productions released until 1993 (the first ten years in our research dataset) are mostly in the hands of Japanese companies—nearly all of which are located in the Kantō and Kansai region of Japan’s main island, Honshu. The most recent period covered by our data, 2005–2015, reveals that many
more companies located in many more places around the world are contributing to the production of the corresponding releases in our sample. In other words, game production for the Japanese market appears to have undergone diversification and internationalization. In contrast to a general notion of “globalization,” however, our data allows us to pinpoint specific areas where the production of games released in Japan is concentrated, namely North America, Europe, and Asia. The empty patches on the map are just as informative as the populated ones: Videogame production for the Japanese market is largely an undertaking of the Global North. At the same time, Japan occupies a central space in this picture, given that a large number of companies with headquarters in Japan appear on this production map. Instead of corroborating or falsifying the globalization hypothesis outlined in the introduction, this approach offers a more detailed view of a highly complex issue that needs further unpacking. We are currently exploring several paths toward that end.

Our experiences show that the process of generating research datasets from online resources involves a series of important decisions that, in turn, inform the analysis and interpretation of the data. Moreover, the process is still a technical challenge, especially when dealing with datasets that involve different languages and do not feature unique identifiers. The complex nature of videogame titles, furthermore, prevented us from using standard string similarity measures and led us to develop our own algorithm for comparing videogame titles. Overall, we managed to identify corresponding entries for about 40 percent of the Media Arts Database entries [of our datasets]. This process was both time-consuming and technically challenging, but it also allowed us to gain a better understanding of the datasets and the specific features of videogame titles in general. In addition, we believe that this algorithm can be expanded and that it can assist videogame entity recognition and deduplication tasks.

Besides these technical challenges, we identified two other key decisions in the research process that influenced the resulting research dataset: the conceptual exploration of the data resources and the sampling strategies that we used to counter biases in our sources. While our data sources describe the same domain, their coverage differs, resulting in a bias in the usable (matched) data. Weighting a sample can minimize the introduction of such biases. With multiple databases covering different, overlapping subsets of videogames, we showed that a stratified sampling approach aids the drawing of a representative sample. More generally, in order to maintain a level of empirical representability, the available data forced us to shift our perspective, from games produced in Japan to games released in Japan.
Furthermore, our investigation of two databases revealed that both offer different views on what exactly a videogame consists of. This allowed us to discuss the results in a more precise way; specifically, we do not talk about games, we talk about local releases of games. Therefore, when combining both datasets, we have to consider the possibility that combined information from editions (MobyGames) and local releases (Media Arts DB) will lead to further questions: for example, when should companies doing local English translation work on a game first released in Japan be considered actors involved in the development of said game?

The links between the datasets also provide the foundation for a comprehensive knowledge base on videogame (meta) data. By linking more data sources and creating a more detailed vocabulary, this process facilitates the iterative construction of a knowledge base that, in turn, can be used to create better research datasets for future research questions. On the other hand, the selection of data sources is informed by both the research question and more technical aspects. It can therefore be seen as a deliberate act of curation and should be reflected as such in the research process.

We must consider the above-mentioned challenges when building each new research dataset from available online data sources: there are no raw data, meaning that besides structural and conceptual differences we must also identify and deal with the biases and curatorial vernacular practices of each data source if we are to build research datasets that are meaningful and reusable. That said, we believe that, with some case-specific adjustments, our approach can be deployed in a similar way in other data-driven research, thus contributing to the range of methods and tools available to any researcher or research community interested in using data stored in heterogeneous online resources.

References

Aarseth, Espen. 2011. “Define Real, Moron! Some Remarks on Game Ontologies.” In Digarec Keynote-Lectures 2009/10, edited by Stephan Günzel, Michael Liebe, and Dieter Mersch, 50–69. DIGAREC Series 6. Potsdam: Potsdam University Press. https://nbn-resolving.org/urn:nbn:de:kobv:517-opus-49780.
Aarseth, Espen, and Gordon Calleja. 2015. “The Word Game: The Ontology of an Undefinable Object.” Proceedings of the 10th International Conference on the Foundations of Digital Games (FDG 2015). [Santa Cruz, CA]: Society for the Advancement of the Science of Digital Games. http://www.fdg2015.org/papers/fdg2015_paper_51.pdf.

Barocas, Solon, Danah Boyd, Sorelle Friedler, and Hanna Wallach. 2017. “Social and Technical Trade-Offs in Data Science.” Big Data 5 (2): 71–72. doi:10.1089/big.2017.29020.stt.

Christen, Peter. 2012. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Berlin: Springer.

Cohen, William W., Pradeep Ravikumar, and Stephen E. Fienberg. 2003. “A Comparison of String Metrics for Matching Names and Records.” Paper presented at KDD Workshop on Data Cleaning and Object Consolidation, KDD03: The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C., August 24–27, 2003.

Consalvo, Mia. 2009. “Convergence and Globalization in the Japanese Videogame Industry.” Cinema Journal 48 (3): 135–41. doi:10.1353/cj.0.0122.

Consalvo, Mia. 2016. Atari to Zelda: Japan’s Videogames in Global Contexts. Cambridge, MA: MIT Press.

Gitelman, Lisa, and Virginia Jackson. 2013. Introduction to ‘Raw Data’ Is an Oxymoron, edited by Lisa Gitelman, 1–14. Cambridge, MA: MIT Press.

Goto-Jones, Chris. 2015. “Playing with Being in Digital Asia: Gamic Orientalism and the Virtual Dōjō.” Asiascape: Digital Asia 2 (1–2): 20–56. doi:10.1163/22142312-12340019.

Harron, Katie, Harvey Goldstein, and Chris Dibben. 2016a. Introduction to Methodological Developments in Data Linkage, edited by Katie Harron, Harvey Goldstein, and Chris Dibben, 1–7. Somerset: Wiley.
Harron, Katie, Harvey Goldstein, and Chris Dibben. 2016b. “Summary.” Chapter 10 in Methodological Developments in Data Linkage, edited by Katie Harron, Harvey Goldstein, and Chris Dibben, 226–32. Somerset: Wiley.

Jett, Jacob, Simone Sacchi, Jin Ha Lee, and Rachel Ivy Clarke. 2016. “A Conceptual Model for Video games and Interactive Media.” Journal of the Association for Information Science and Technology 67 (3): 505–17. doi:10.1002/asi.23409.

Jiang, Yu, Can Lin, Weiyi Meng, Clement Yu, Aaron M. Cohen, and Neil R. Smalheiser. 2014. “Rule-Based Deduplication of Article Records from Bibliographic Databases.” Database 2014 (article ID bat086). doi:10.1093/database/bat086.

Kalashnikov, Dmitri V., and Sharad Mehrotra. 2006. “Domain-Independent Data Cleaning via Analysis of Entity-Relationship Graph.” ACM Transactions on Database Systems 31 (2): 716–67. doi:10.1145/1138394.1138401.

Kerr, Aphra. 2017. Global Games: Production, Circulation and Policy in the Networked Era. New York: Routledge.

Köpcke, Hanna, and Erhard Rahm. 2010. “Frameworks for Entity Matching: A Comparison.” Data & Knowledge Engineering 69 (2): 197–210. doi:10.1016/j.datak.2009.10.003.

Lee, Jin Ha, Rachel Ivy Clarke, Simone Sacchi, and Jacob Jett. 2014. “Relationships among Video Games: Existing Standards and New Definitions.” In Proceedings of the American Society for Information Science and Technology 51 (1). doi:10.1002/meet.2014.14505101035.

Liu, Jialu, Kin Hou Lei, Jeffery Yufei Liu, Chi Wang, and Jiawei Han. 2013. “Ranking-based Name Matching for Author Disambiguation in Bibliographic Data.” In KDD Cup ’13: Proceedings of the 2013 KDD Cup 2013 Workshop, article no. 8. doi:10.1145/2517288.2517296.

Manovich, Lev. 2017. “Cultural Data: Possibilities and Limitations of Working with Historical Cultural Data.” In Museum and Archive on the Move: Changing Cultural Institutions in the Digital Era, edited by Oliver Grau with Wendy Coones and Viola Rühse, 259–76. Berlin: De Gruyter.

Macias-Garcia, Miguel A., Victor J. Sosa-Sosa, and Ivan Lopez-Arevalo. 2009. “A Generic Approach for Data Integration using RDF, OWL and XML.” In: Artificial
Intelligence and Applications. Complementary Proceedings of MICAI 2009, 8th Mexican International Conference on Artificial Intelligence. http://www.micai.org/2009/proceedings/complementary/cd/ws-mldm/215/20091105_MiguelMacias_Micai09_workshop_ML_and_DM.pdf.

MobyGames. n.d. “The MobyGames Standards and Practices,” section 1.2.3, “New Entry.” Updated August 24, 2019. https://www.mobygames.com/info/standards#New_Entry.

Mühleder, Peter, Florian Rämisch, and Tracy Hoffmann. 2019. “Interlinking Dataset of the Video Game Databases MobyGames and MediaArt Database.” Data set. Zenodo. doi:10.5281/zenodo.3356507.

Nicoll, Benjamin. 2017. “Sega Saturn Fan Sites and the Vernacular Curation of Videogame History.” In Fans and Videogames: Histories, Fandom, Archives, edited by Melanie Swalwell, Helen Stuckey, and Angela Ndalianis, 180–96. New York: Routledge.

Winkler, William E. 2016. “Probabilistic Linkage.” In Methodological Developments in Data Linkage, edited by Katie Harron, Harvey Goldstein, and Chris Dibben, 8–35. Somerset: Wiley.

Wolf, Mark J. P. 2015. Introduction to Videogames around the World, edited by Mark J. P. Wolf, 1–16. Cambridge, MA: MIT Press.