Solon Scott
02/26/2026
LIS 545 – Data Curation
Term Project – Final Report

Data and Metadata Profile

Source:
https://figshare.com/articles/dataset/Encuesta_Gamblificaci_n_juventud_y_videojuegos/30467558?file=60987490

I've found an interesting data package on Figshare by a team of researchers from Spain investigating the effects of gambling mechanics within videogames on the emotional wellbeing of young people. This is very new data that was posted two weeks ago on January 13th, 2026, and it doesn't come with any published work. Unless I find something to suggest otherwise, I am assuming this is the finished results as they've submitted this Figshare project to be the central DOI to this work. Funding for this work comes from the Spanish Government's Ministry of Social Rights for their 2030 Agenda, which means the results from this research could have civic repercussions within Spain. There are also three authors for this data, Aranda, Fernández-de-Castro, and Sáenz-Leandro, they seem to be affiliated with UOC Open University of Catalonia in Barcelona, Spain based on the associated email addresses. Within this package is five items: a survey (ES), the results of the survey (N/A), a codebook for the results (ENG), a readme file (ENG) to contextualize this package regardless of whether it is hosted on Figshare or not, and a list of the algorithms that were run on the data in R to replicate findings from the research (ES). The 35-question survey is in a PDF format that seems to be adapted from a survey hosting webpage. The survey also splits into two sub-sections for folks who answer whether they consider themselves "Gamers" or "Non-gamers". The results feature 1000 participants and are in a plain CSV format that are coded to work well with most spreadsheet applications. The codebook is in XLSX format which is proprietary to Microsoft Excel but seems to adapt well for other spreadsheet applications like Google Sheets. The readme is a plain TXT file which most online repositories like Github should be able to use as a frontend, and the algorithms are published as a .R format which is useable for data analysis programs like R but they can also be read in plaintext with minimal metadata loss. The data package has a CC BY 4.0 Creative Commons license allowing anyone to share and adapt the work as long as appropriate credit is given to the authors and no additional restrictions are applied.

The codebook provided as a separate file from the data is a substantial record of human-readable (not machine-readable) instructions to go along with the primary data. I find it interesting that the metadata and readme documents are in English, because even without

knowing any Spanish, I'm able to fully access this data and process it further. It makes me wonder if English is a defacto language for metadata in certain regions or if it is better for access by software that may be developed in primarily English-speaking areas. The codebook features a list of all variables, their labels, what type of data structure it is in, the value range each variable can reach, what may constitute missing values, semantic categories for items "per block", and "routing notes" to show which survey questions where answered by which all participants and when participants were routed to other questions. This codebook successfully encapsulates the data provided and is well formatted for common codebook guidelines but does not seem to be structured according to any metadata standards that would allow for machine-ready transformative use.

The Readme could be more robust to help contextualize the data, especially for describing what state the data is in within the overall project or how it could be extended towards being machine-readable. In the current format the metadata are in now it would take a researcher some effort to re-prepare this data for being workable in a data analysis program. If improved to standards such as DDI's it would ensure data are machine-actionable, consistent, and FAIR. My other suggestion is to expand the findability of this data by looking for broader keywords than 'Gamblification' as used on Figshare. These could also be added to the Readme in case this project needs to ported to another repository.

This is a very new dataset and is not yet currently listed with any broader publications on the data, but it has a strong base to work from and after review it seems ready to provide strong value on how young people interact with games that feature luck-based reward systems. All that is needed now is to get it to be made machine-readable and archivable.

Repository Profile

Repository Profile on Greo Dataverse - https://borealisdata.ca/dataverse/Greo

The dataset I am working with adeptly starts with the word "Gamblification" even though it is not a very strong keyword but it gets across an intense academic reaction that can help put it into specific groups. From exploring re3, it seems repositories are one of those kinds of extended groups that can identify a datasets overall 'field'. Looking for videogames or gaming won't lead to an urgent end and thus no specific repositories show up. Youth psychology becomes a very broad field where way too many repositories show up, but 'gambling' leads to exactly one repository: Greo Dataverse, and that feels like a really good sign that the wider academic universe knows very precisely what to do with this sort of data.

Greo is a Canada-based repository, open for submissions of tabular data, textual documents, digital images, and even some video formats but with a limit of 3GB. It seems like anyone can upload as long as it is within the field of "preventing or reducing harm related to gambling, gaming, technology use, and substance use." Because Greo is not a very large repository, these seem more like systemic limitations set by the Dataverse software being used (Borealis), as Greo seem very open and available to working with groups to deposit larger amounts of data. The primary guidance for submissions are two things:

- Simple rules: Must be Final Draft, digital, and without identifying information
- Descriptive metadata: As much as is reasonably possible to support understanding, discovery, and reuse (Title, Authors, Description, Date, Methodology, Funding, Citations, etc.)

Once again, this also feels very flexible as any submission will see direct human assistance and consulting once the process starts. So if pieces are missing or qualifications are not being met, there will be a person who can help get datasets into working order for the repository. They even offer services to help develop codebooks for datasets amongst many other services. Based on RE3's information, Greo features four metadata standards: DataCite, DDI, Dublin Core, and OAI-ORE which seems like it would be enough to cover most of the metadata needs they could run into with the various datasets and artifacts that are currently able to be submitted.

Login is not required to access and download data that is freely accessible, but an account can be freely created in order to contact dataset managers and get access. Users can also use institutional login credentials, which will better connect them with material that may require elevated permission to be accessed. The Borealis system’s only access method is currently direct file download – so there isn’t any API access, unfortunately. But this is fine for how few files are currently managed by Greo (151 as part of 31 datasets total). Greo's datasets feature DataCite's metadata standards to manage DOIs and associate metadata, which had their most recent update in November of 2025. Greo does not add anything specific to their DIP beyond what is in a specific dataset package but it does intervene when needed to supply their Terms of Data Use before adding researchers to their list of permitted data users as well as provide an uploading template to be followed when submitting.

Overall, I believe Greo could be a good fit for our Gamblification dataset and its dedicated and accessible team could help make sure it stays available on the repository for a long time. They are a veteran team that have maintained this repo for a long time and have all of the bells and whistles I'd expect for a repo of this size. The only foreseeable drawbacks is

that nearly all of their datasets are focused on Canadian gambling and even though the codebook and dataset are translated to English, they may have some issues supporting the parts of the research that are in Spanish. But maybe that diversity is just what they need?

Recommended Data Citation

Aranda, D., Fernández-de-Castro, P.; & Sáenz-Leandro, R. (2025). Gamblification, Youth and Digital Games: A Dataset on Gaming Practices and Emotional Impacts Among Spanish Young People in Spain. [Figshare]. Univseritat Oberta de Catalunya. https://doi.org/10.6084/m9.figshare.30467558.v2

Considerations for long-term preservation

These data have multiple contexts added to allow for long-term preservation. The survey is recreated in PDF form to give textual context for the data, the data comes with a codebook to give the data context, the replication data is saved in a plaintext format (R) that can provide instructions on how data was manipulated in a way that can be replicated, and the README provided carries links between the DOI, what the dataset means, and instructions for usage and sharing. The only thing missing is that the README does not mention the replication data or what it is used for.

Copyright License Statement

The current license for this data is Creative Commons Attribution 4.0 International which allows users to Share and Adapt with Attribution and it makes sense to extend this license for our own uploading purposes.

Human Subject Considerations Statement

The questionnaire asks for some personal data from human subjects such as age and how much money they spend on games. This data has been anonymized and coded with a separate codebook with measured ethical judgement and to make it more machine-readable.