Helping Users Find New Video Games Using Scraped Data

Contents

Introduction

When searching for new video games, it can be hard for uses to find games that they might be interested in and it can be hard for businesses to reach new users. To help with this, using publicly data from the internet, a database could be built that, with a user interface such as a web application, could help users find games they are interested in that are similar to ones they already like. With this in mind, this project aims to create a web application that is able to help users find video games they may be interested in through the use of a tagging system.

To start, this document will conduct a review of the literature regarding this topic, specifically in how it applies to users and businesses. Then, a consideration will be made concerning the issues this project might have before going over its methodology. Afte defining the methodology, the data itself will be considered and outlined followed by the development of the project's artefacts and what results from it.

Design and Planning

Why a Tagging Website?

When searching for video games to play, users have the options of either going to a physical shop and buying first hand or second hand or going online and buying from a vast number of digital stores. For both options, users will be faced with an overwhelming number of video games to choose from. Castillo (2013) states that this a problem due to people ignoring most options on a shelf after a certain amount and focusing on a small selection instead, specifically a selection of 3 items, while also taking longer to choose. Castillo (2013) also talks about how, when someone knows or has expertise about the items or options in a choice, they will be able to give a more confident choice than when they know little about the items or options.

It is not just the number of choices a user has that makes choosing more difficult, but it could also be the increasing cost of video games (Samuli, 2022) through increasingly predatory practices such as:

- Microtransactions which lock in game content behind real world payment (Samuli, 2022) which come in a large variety of implementations such as “loot boxes” (Samuli, 2022) (Raneri et al, 2022) which potentially put those with gambling disorders at great risk (Raneri et al, 2022).
- Subscription models which only allow a user to play a video game for a short period of time before needing to pay to continue accessing it (Samuli, 2022).

These practices make games more expensive than their initial price tag, which may, along with the number of options to choose from, make it harder for a user to decide on a video game. One more reason for difficulty in choice might also be judging a video game’s cover to represent the video game overall, however, Iwana et al (2016) found that, when trying to create a model to predict a book’s genre by its cover, a lot of books had ambiguous features (or few visuals) leading to incorrect predictions. This might be the case for video games as well.

Overall, the difficulty in choice is the main reason for this project. Castillo (2013) recommends categorisation and restrictions to make choices easier and with more confidence. Assumedly, they imply that categorising helps users make more informed decisions due to knowing what the choices involve through said categorising, and assuming this is true, a website that gives users more information on video games and allows them to search for similar video games through categories (tags and genres) could help them choose new video games to play.

What is Scraping? Why Use It?

Scraping is, as described by Smith (2019) and Zhao (2017), a way to use the internet to collection public information for a specific purpose such as data analysis. It allows an easier way to collect information from a public source without needing to request it (Smith 2019) This can be done by a human manually or a bot automatically (Zhao, 2017). An example of this given by Smith (2019) is for the indexing of a search engine by having a scraping program go from website to website, using links on each, to collect the information on each page and allow each page to be search by comparing search requests to the contents of each page found while scraping. Smith (2019) also goes on to state the usefulness of a web scraper in building a custom dataset in the common event where the required data is not already available.

Xiao (2021) raises a problem that comes with scraping, which is that of data privacy due to a web scraper being able to collect data posted publicly by users online without their consent. While the project will not collect any personal information from any users, data privacy will still be considered by investigating the GDPR and its UK equivalent. However, this problem is not only limited to users' personal data, but companies' data as well along with scrapers potentially violating a company's terms of service (Zhao, 2017). However, Zhao (2017) argues that it is difficult to prove the data's copyright due to "only a specific section of the data [being] legally protected", while arguing that terms of service fall into a "gray area". Nonetheless, all of the data scraped in this project should have its source saved and, in the website, linked to in order to show that the data is from the target website and not from another source.

Overall, the main reason for using a scraping program to collect data on video games is because there are a lot of video games and a lot of information regarding them. Collecting this data manually would take a great amount of time and analysing it even more so. Scraping would allow for the automation of this data collection while allowing for a large amount of this data to be collected in a reasonable time.

Legal and Security Issues

GDPR

The data scrapped from the target websites will not contain any personal information, however, the website will to a very minor degree. This information will be limited to a user's email address and will be used for the sole purpose of allowing a user to create an account on the website; no other information will be taken. Because of this, both the GDPR and the UK's own data protection laws will be described here and implemented into the website.

As per the GDPR, this project needs to:

- Use data in a lawful, fair and transparent manner.
- Use data for only the purposes stated.
- Use only the necessary data needed for the stated purpose.
- Keep collected data accurate and up to date.
- Store data for only as long as it needs to be for the specified purpose.
- The data controller (me) must be held accountable for stored data.

This project also needs to follow users' rights, which the GDPR describes as the rights to:

- Be informed about data's usage.
- Have access to stored data.
- Have data be rectified.
- Have data erased upon request.
- Have data's processing be restricted.
- Have data be portable to be used in other services.
- Object to data's processing.

(GDPR, n.d.)

While similar, the UK's data protection legislation adds the following:

- Data must be handled securely to prevent against unauthorised (or illegal) processing, access, loss, destruction or damage.
- More sensitive information such as race, ethnic background, political opinions, religion, sexual orientation, etc., have stronger legal protections.

(United Kingdom, n.d.)

To implement this, we should inform users about their email's usage, allow them to modify it at any time and allow them to delete their accounts whenever they request to.

Computer Misuse Act 1990

Due to scraping taking data from websites, the scraping program must be made in a way such that it cannot take data from an unauthorised source (CPS, 2006). Websites that would require a login or further credentials to access data should not be used. Steam is a good target because, while you need to login to buy games, you do not need to sign in to access its game pages.

The Copyright, Designs and Patents Act 1988

Once a person creates something, they have copyright over their creation for a number of years depending on what kind of creation it is (for example, literary, dramatic, musical or artistic works have a copyright of 70 years, while sound recordings have 50 years instead). This copyright gives the person the right to control how their work is used and allows them to object to distortions of their work due to being identified as the author (UKCS, n.d.).

Because of this, the scraping project should, again, focus only on public data that has been put up with the author's approval. Video games viewable on public websites such as Steam implies the author's consent for their works to be viewable.

STRIDE

STRIDE modelling is a form of security modelling for software developers that revolves around predicting potential threats during the design process of a program and attempting to mitigate them (Landuyt, 2022). This model stands for:

- Spoofing: A spoofing (faking) the identity of another user.
- Tampering: Unauthorised modification of data weather intentional or accidental.
- Repudiation: A user claiming that they did not commit an action with no way for a system to prove that they did (I.E: logging).
- Information disclosure: Unauthorised access and reading of data.
- Denial of service: Denial of Service or Distributed Denial of Service attacks.
- Elevation of privilege: A user without administrative access to a system gains administrative access and gains the ability to compromise the system.

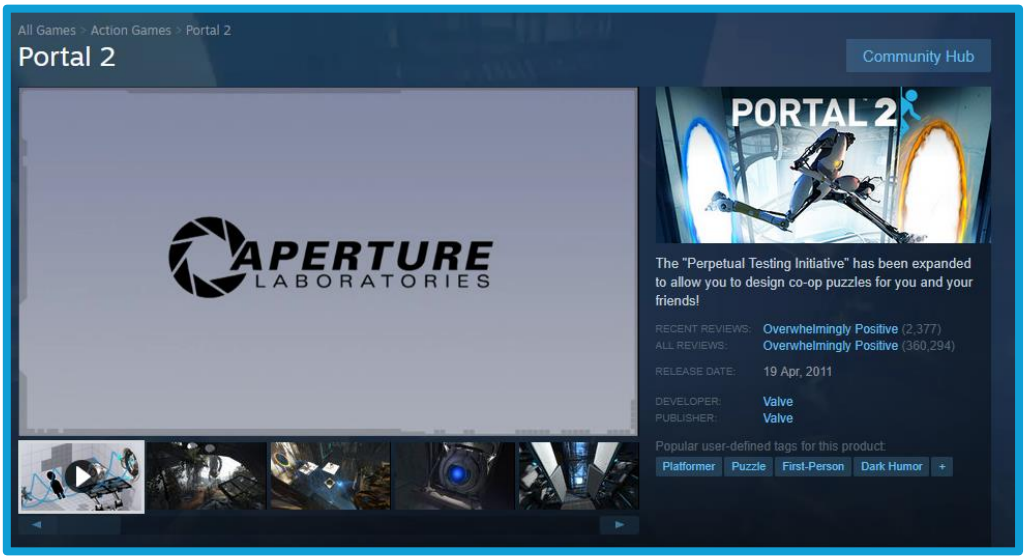
(Microsoft Learn, 2009) (Conklin et al, n.d.).

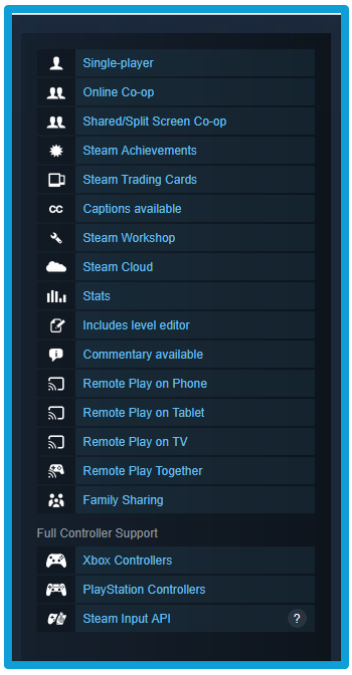
To help against these threats, the following mitigation should be done for the website:

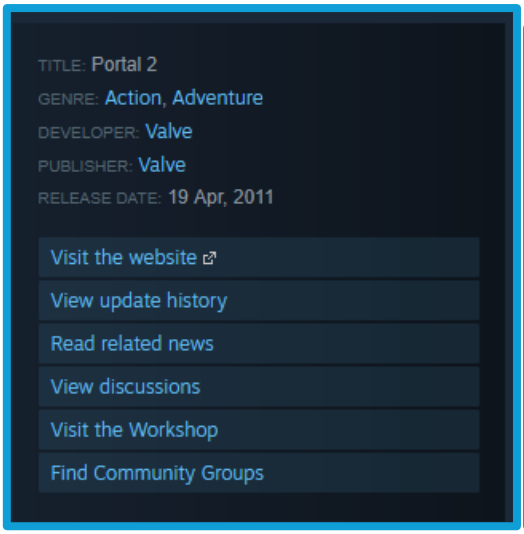
Threat	**Mitigation**	**Meaning**
Spoofing	User Authentication	Ensure the user is who they say they are through techniques such as two factor authentication.
Tampering	Data Integrity	Use encryption and hashes to protect sensitive data. Ensure sensitive data requires authorisation to modify.
Repudiation	Non-Repudiation	All actions by users should be logged by the system to prove a user's actions.
Information disclosure	Confidentiality	Only those with authorisation should be allowed to read data. Data should be encrypted.
Denial of Service	Availability	Access to the service should be throttled should users repeatedly send requests to it.
Elevation of privilege	Authorisation	Ensure users are only given the permissions needed to perform their role.

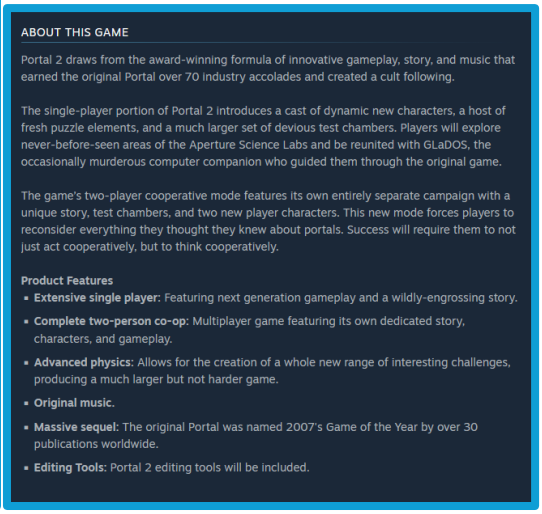
Ad-Hoc Data Issue

Due to this project using scraping, it runs the risk of collecting data that is not entirely accurate and data that might contain biases or ethical issues. However, the main target for this program is the online video game distributor known as Steam. However, when publishing a video game on Steam through the user of “Steamworks”, the author must build their video game’s store page. The only part Steam does in the creation of this store page is that they review it after submission in order to confirm that it is working and not harmful (Steamworks, n.d.). Because this information is submitted by the author themselves, this project assumes that each page contains mostly accurate information about the video games themselves. Missing or erroneous information will be fixed after scraping through normalisation. Below are example images of the store page on Steam for the video game Portal 2 to show this from Steam (n.d.):

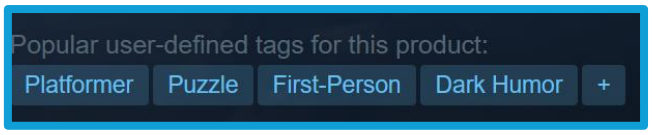






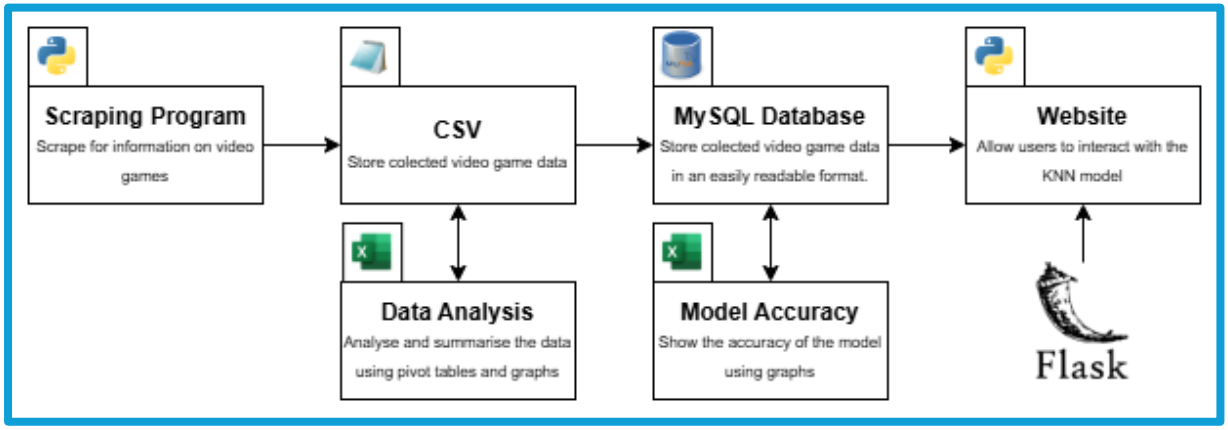


There is one part of these pages, however, that is not decided upon by the author. This is the video game tags which are described on the video game's page as "Popular user-defined tags" (Steam, n.d.). This can be seen below:



Because of this, while this information will still be collected, it would be best for there to be some sort of review process for these tags to make sure that they match the game they describe.

Methodology



Overall, the main goal of this project is to create a website that allows for users to search for video games similar to ones they already like. To achieve this, this project will be split into the following parts which can be seen in the above diagram:

1. Scraping program: As previously described, this program will go over the target website (in this case, Steam), and collect data about various video games. This should be limited to low number of video games to avoid affecting Steam with high amounts of traffic and to avoid the possibility of running out of storage or the program taking too long.
2. CSV: The data is then saved using the comma separated values format which will then be reviewed and normalised using Excel or Python.
3. MySQL database: The normalised data will be loaded into a MySQL database using Python. This database will work alongside the final part:
4. Website: This will be what the user will use to search for video games using data from the MySQL database.

Once the website has been created, the final step will be to review the scraping program, the database and the website to see how well they meet the goals the project wishes to achieve and to determine what can be done to all parts to improve them for future use.

Data Plan

Scraping

For the website to work, it needs a database full of information about various video games and that information needs to describe those video games in different ways to differentiate all of them. To get this data, as previously described, a scraping program will used that will be programmed in Python.

To do this, several libraries could be used:

- Requests: This library allows Python to send HTTP requests to targeted web pages. More specifically for this project, the HTML can be requested which contains the information that would be displayed on a webpage (Pypi, n.d., -a).
- Beautifulsoup: This library is a HTML (and XML) parser that allows for easy manipulation and searching of a HTML tree (Pypi, n.d., -b).
- Selenium: This library allows Python to take control of a browser (Pypi, n.d. -c) in a programmatic way which allows for web scraping like the requests library but in a browser instead of just a HTML file (GeeksForGeeks, 2025).
- Schedule: This library allows for the scheduling of Python functions allowing for Python functions to be run periodically without a human needing to run it again (Pypi, n.d. -d).

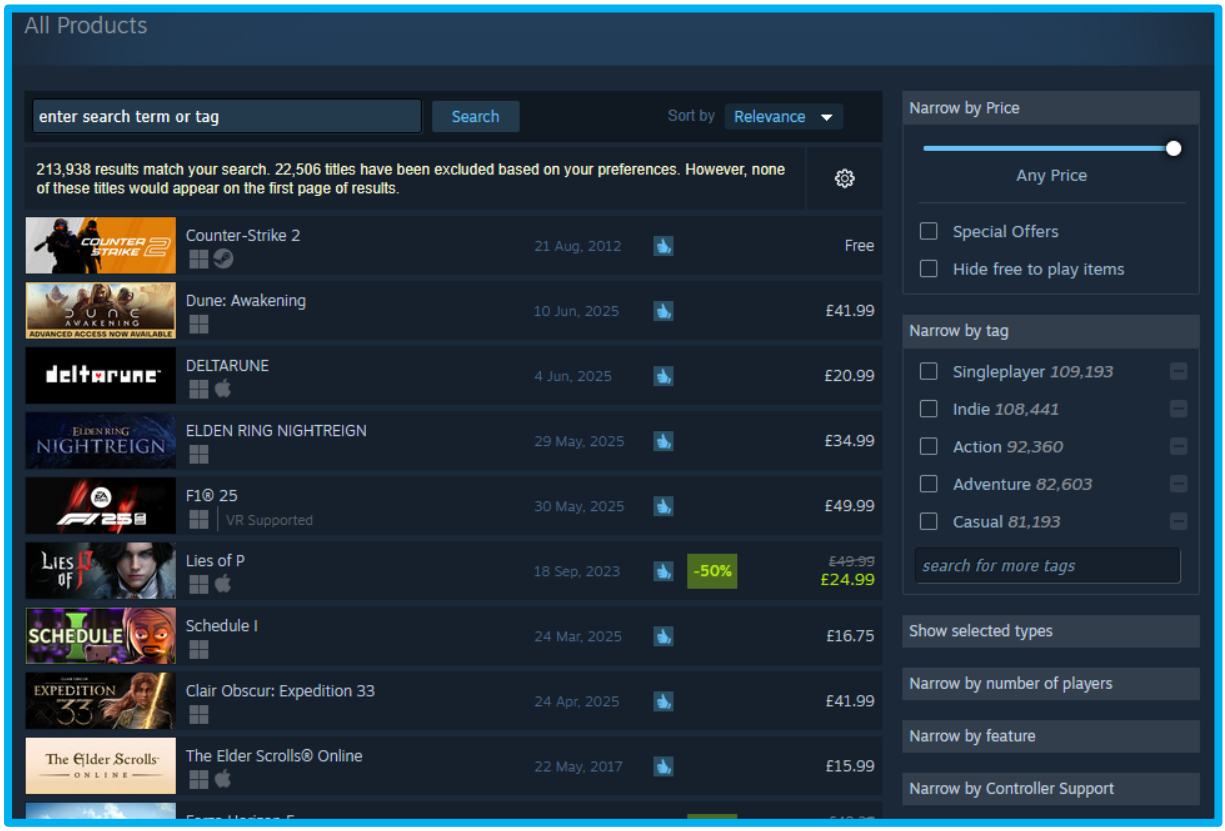
(GeeksForGeeks, 2025)

The next part of this project is to learn how scraping works which will involve deciding which of these libraries (or others) should be used.

For the scraping program itself, the target website must have a page that lists *all* of its games. Using this page, the following should be done for each game in the list:

1. Access the link to the game's page.
2. Systematically go through the game page's HTML to get information the video game.
3. Save the video game's information gathered by appending it to a CSV file.
4. Return to the video game list and start again on the next game in the list.

Target Website



For the scraping program to work, it needs a target. The target in this case is the previously mentioned online video game retailer, Steam.

Steam has hundreds of thousands of video games available to collect data from (the above screenshot shows 225330 games for an empty search). As stated previously, all data on each video game page is submitted by the developers themselves and this page in the screenshot meets the previously mentioned requirement of there being a page with a list of all games which will make the scraping program easier.

This page also gives the scraping program a lot of ways to sort the list if needed.

Saving Format (.CSV)

A comma separated values (Creativyst Software, n.d.) (Microsoft Support, n.d.) file is a filetype that stores tables in plaintext with each row and column being separated by commas (Microsoft Support, n.d.). Both Creativyst Software (n.d.) and Microsoft (n.d.) state that CSV files make it easier to share information between multiple programs with Microsoft allowing CSV files to be imported into their Outlook software.

In Python, CSV files can be opened in the same way as text files and, with the CSV library, the CSV read can allow Python to handle files as lists of strings with each list representing a row (Python Documentation, n.d.).

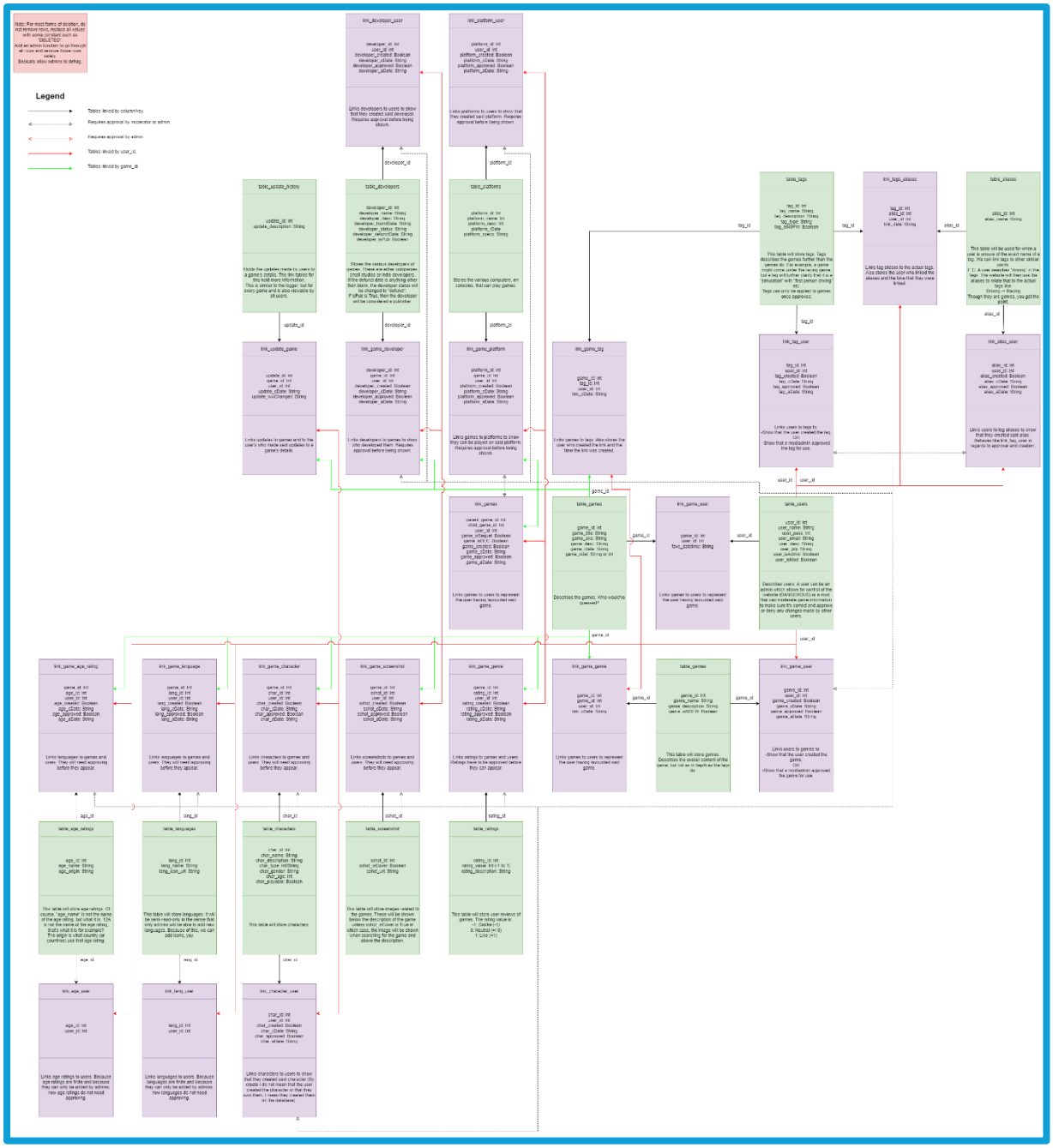
Overall, this ease of use is the reason why this project will be using CSV files. They allow the easy saving of scraped data, which can then be loaded into Excel for normalisation. This normalised data can then be saved as another CSV file that can easily be loaded into the website using Python again.

Database (MySQL)

MySQL is an open-source database management system that allows for a large amount of data to be stored and managed (MySQL, n.d.). MySQL is relational meaning that each database has multiple tables that have relationships defined by the programmer (MySQL, n.d.). As shown in its name, MySQL is programmed with the structured query language (SQL) which is a standardised language for handling databases (MySQL, n.d.).

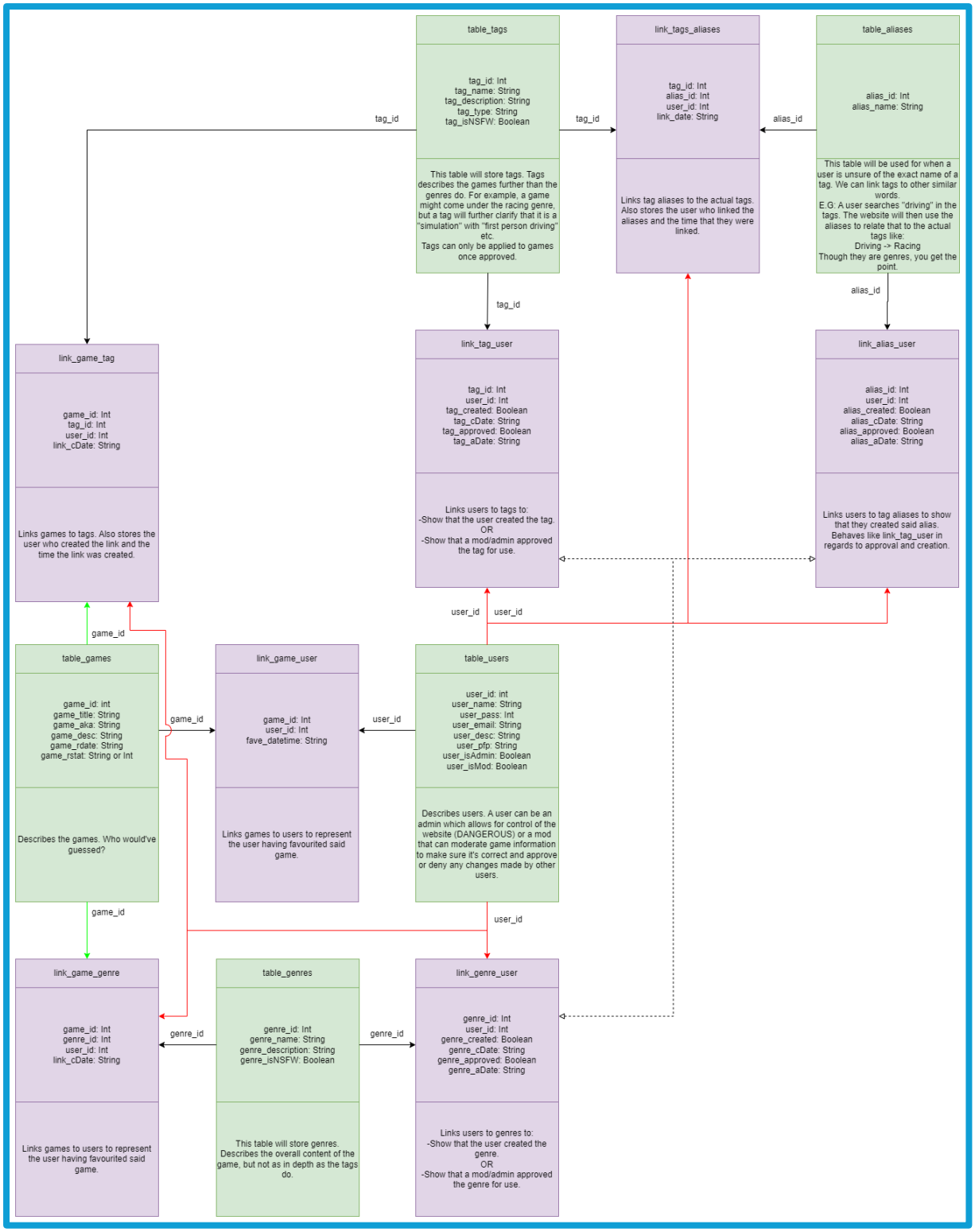
This project will use a MySQL database to store the data scraped from the internet after said data is loaded into the database from the CSV files. The reason for using MySQL is the ease in which Python can communicate with the database to store and receive data using the MySQL Connector Python library. This library allows Python to communicate with a MySQL database through an API which allows for the database to be connected to and, through the use of a cursor, allows for the execution of SQL commands on the database (Pypi, n.d. -e).

(Continued…)



Above is a plan for the database. In it, the green boxes are regular tables that define different bits of data. The purple boxes, however, are link tables which describe how the data in two or more tables are linked. At the most basic level, the link tables hold the ids of the columns that are linked. For example, a user in the user table linked to their review in the review table will have a column in the link_user_review table that holds the user's ID and their review's ID.

(Continued…)



Here is a clearer diagram that shows only the games, users tags and tag aliases tables.

Development

Scraping Program

MySQL Database

Website

Results/Findings

Scraping Program

MySQL Database

Website

Conclusion

References

- Castillo, M. (2013, Jun). Overwhelmed by Choices. American Journal of Neuroradiology : AJNR, 34(6), 1111–1112. https://doi.org/10.3174/ajnr.A3274
- Conklin, L., Drake, V., Strittmatter, Sven., Braiterman, Z., & Shostack, A. (n.d.). Threat Modeling Process. OWASP. https://owasp.org/www-community/Threat_Modeling_Process
- CPS. (2023, August 3rd). Computer Misuse Act. The Code for Crown Prosecutors. https://www.cps.gov.uk/legal-guidance/computer-misuse-act
- CPS. (2024, March 11th, b). Fraud Act 2006. The Code for Crown Prosecutors. https://www.cps.gov.uk/legal-guidance/fraud-act-2006
- GDPR. (n.d.). What is GDPR, the EU's new data protection law? GDPR. https://gdpr.eu/what-is-gdpr/
- Creativyst Software. (n.d.). How To: The Comma Separated Value (CSV) File Format. Creativyst Software. https://www.creativyst.com/Doc/Articles/CSV/CSV01.shtml
- GeeksForGeeks. (2025). Python Web Scraping Tutorial. GeeksForGeeks. https://www.geeksforgeeks.org/python-web-scraping-tutorial/
- Iwanna, B., Rizvi, S., Ahmed, S., Dengel, A., & Uchida, S. Judging a Book By its Cover. Cornell University. https://doi.org/10.48550/arXiv.1610.09204
- Landuyt, D., & Joosen, W. (2022). A descriptive study of assumptions in STRIDE security threat modelling. Software and Systems Modelling, 21(6), 2311-2328. https://doi.org/10.1007/s10270-021-00941-7
- Microsoft Learn. (2009). The STRIDE Threat Model. Microsoft Learn. https://learn.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v=cs.20)?redirectedfrom=MSDN
- Microsoft Support. (n.d.). Create or edit .csv files to import into Outlook. Microsoft Support. https://support.microsoft.com/en-gb/office/create-or-edit-csv-files-to-import-into-outlook-4518d70d-8fe9-46ad-94fa-1494247193c7
- MySQL. (n.d.). 1.2.1 What is MySQL? MySQL Documentation. https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html
- Pypi. (n.d. -a). requests 2.32.3. Pypi. https://pypi.org/project/requests/
- Pypi. (n.d. -b). beautifulsoup4 4.13.4. Pypi. https://pypi.org/project/beautifulsoup4/
- Pypi. (n.d. -c). selenium 4.33.0. Pypi. https://pypi.org/project/selenium/
- Pypi. (n.d. -d). schedule 1.2.2. Pypi. https://pypi.org/project/schedule/
- Python Documentation. (n.d.). csv – CSV File Reading and Writing. Python. https://docs.python.org/3/library/csv.html
- Pypi. (n.d. -e). mysql-connector-python 9.3.0. Pypi. https://pypi.org/project/mysql-connector-python/
- Raneri, P., Montag, C., Rozgonjuk, D., Satel, J., & Pontes, H. (2022, June). The role of microtransactions in Internet Gaming Disorder and Gambling

Disorder: A preregistered systematic review. Addictive Behaviors Reports. 15, 100415. https://doi.org/10.1016/j.abrep.2022.100415

- Samuli, K. (2022, January 19th). Pricing economics of video games: a panel data study on the effects of versioning on revenue. University of Oulu Repository. https://urn.fi/URN:NBN:fi:oulu-202201191079
- Smith, Vincent. (2019). Go Web Scraping Quick Start Guide (1st Edition). Packt Publishing.
- Steam. (n.d.). Portal 2. Steam. https://store.steampowered.com/app/620/Portal_2/
- Steamworks. (n.d.). Steam Direct (Joining The Steamworks Distribution Program). Steam. https://partner.steamgames.com/steamdirect
- United Kingdom. (n.d.). Data protection (The UK's data protection legislation). United Kingdom. https://www.gov.uk/data-protection
- UKCS. (n.d.). UK copyright law: An introduction (The Copyright, Designs and Patents Act 1988). The UK Copyright Service. https://copyrightservice.co.uk/copyright/uk_law_summary
- Xiao, G. (2021). BAD BOTS: REGULATING THE SCRAPING OF PUBLIC PERSONAL INFORMATION. Harvard Journal of Law & Technology, 34 (2), 701.
- Zhao, B. (2017). Web Scraping. Encyclopedia of Big Data. https://link.springer.com/rwe/10.1007/978-3-319-32001-4_483-1 ALT: https://www.researchgate.net/profile/Bo-Zhao-3/publication/317177787_Web_Scraping/links/5c293f85a6fdccfc7073192f/Web-Scraping.pdf
-