Katy Granath

LIS 545

Mayernik

01.27.2023

[Canvas Link to Description](https://canvas.uw.edu/courses/1620201/files?preview=100833544)

[Canvas link to submit](https://canvas.uw.edu/courses/1620201/assignments/8056397)

**Data and Metadata Profile**

**The Data**

The data are two CSV files made publicly available by Archive of Our Own (AO3) about the archive’s works and tags as of March 21, 2021. There are no restrictions on the use of the data.

A *work* is a digital document published by a creator and stored and accessed in the web archive AO3. Some works published on AO3 are not written works, but are linked video files, art, or what are called “podfics”, which are audio files like a podcast. These are identifiable in the data as works with a zero (0) wordcount. The vast majority of works on AO3 are literary works.

A *tag* is metadata, an label added by a creator about their document to indicate its age rating, fandom, and other relationships. Tags are utilized by AO3 users to search and filter works. While creators can add any combination of letters or words and create a tag for their work, there is a formal system of “canonical” tags, which are organized and curated by dedicated staff (AO3, 2022).

The 968.1 MB CSV file about works is a list of 1,048,575 work titles, as well as their creation date, language, word count, if restricted or not, if complete or not, and associated tag IDs. The second CSV (581.5 MB) is a list of 1,048,575 tag IDs, and their associated tag type (e.g. Warning, Fandom, Relationship), tag name, if canonical or not, an approximate number of uses, and its merger ID (the tag’s canonical equivalent, if there is one). Tag names are redacted if they're non-canonical and have less than 5 uses.

Software that can interpret a CSV file is necessary. User comments on the AO3 post suggest that Excel may not be able to handle the sheer volume of data in the files, so specialized software may be necessary, or specialized hardware may be necessary, for example a computer may need to meet a threshold of memory or RAM to utilize the file. For example, when I open either of the CSV files using Excel on my Macbook, I receive a message from Excel warning me, “This data set is too large for the Excel grid. If you save this workbook, you'll lose data that wasn't loaded.” After opening the file in Excel, a highlighted message at the top appears, reading, “Possible Data Loss: some features might be lost if you save this workbook in the CSV format. To preserve these features, save it in an Excel file format.” The warning messages disappear when the data are saved in an Excel file format. The identical number of rows in both sheets suggests that Excel is not displaying all of the data in this dataset.

The metadata is fairly limited. It does not include other attributes of works, like number of bookmarks, number of times added to a collection, number of page visits, or number of “kudos” (this is akin to a Twitter “like”). It also does not distinguish between date created and date published. It does not include metadata about creators - for example how many subscribers they have, or how many other works they have published. There is no explicit indication that the metadata follows a standardized structure. The metadata of a work’s language uses a standard abbreviation for languages. The dates are in standardized U.S. date formats (M/DD/YY). AO3 also maintains a robust online repository of resources that define terms, and describe their methods of archiving works and curating tags.

**Stakeholders**

The key stakeholders for the data are three main groups:

* Archive of Our Own staff - the archivists, web content managers, and lawyers, who run both the website and the organization;
* Creators - the creators (most often authors) whose work is the listed data, and who are responsible for tagging;
* Readers - the readers who interact with the work the writers have produced.

These groups are not distinct, mutually-exclusive groups. They should be seen as a Venn diagram that is nearly a circle. Few creators have not enjoyed the work of others, and many readers have tried their hand at writing. The AO3 staff are often actively involved in the fan community, not just as archivists, but as creators or readers.

**Use of The Data**

When AO3 shared the data files on March 21, 2021, they asked for feedback from community members who had used the data. Approximately one month later comments on the post were closed, which disallows further commenting, precluding further knowledge from this source about whether the data has been used. I was not able to find a way to locate who had used this data and for what, outside of the limited comment section on the original March 21 post. In the comments, less than ten users reported using the data. Five users said that they were going to use the data for university-level projects, from data science and analytics, to dissertations. Three other users shared links to githubs and Tableau to show what they had done with the data. One other said that they had already published research with similar data, but that they would use the AO3 data in the future. SeeAppendix A for the list of comments by users who reported they utilized the data.

It’s safe to say that Archive of Our Own (AO3) is wildly popular. Every day the site attracts tens of millions of page views, and tens of thousands of comments (AO3, 2020). The record for the longest novel in the English language has been broken time and again by AO3 fan-fiction authors publishing stories millions upon millions of words long (Castello, 2022). Yet the search phrase “archive of our own” returns only 40 mentions in peer-reviewed journals in the University of Washington library database, only six of which are not from the last five years. Peer-reviewed analysis of fanfiction and its community, if not entirely nascent, is still in the early stages of development.

Because the data that AO3 published can also be acquired by anyone using a method called data scraping, it is possible that there are published works that use similar data, but not this specific dataset. For example, Price and Robinson’s 2021 article, “Tag analysis as a tool for investigating information behaviour: comparing fan-tagging on Tumblr, Archive of Our Own and Etsy”. Other articles about fanfiction may use different data acquired from AO3. For example, Noppe et al’s “What if academic publishing worked like fan publishing?” (2022), or Johnson’s “Fan Fiction Metadata Creation and Utilization within Fan Fiction Archives: Three Primary Models” (2014).

*The following resources may be useful to inform your assessment and response:*

*Research Data Alliance (RDA) Metadata Directory, http://rd-alliance.github.io/metadata-*

*directory/*

*DataOne Best Practices Primer,* [*https://dataoneorg.github.io/Education/bestpractices/*](https://dataoneorg.github.io/Education/bestpractices/)

**Appendix A:**

User comments of use on the Archive of Our Own data dump post on March 21, 2021

1. User [Androxys](https://archiveofourown.org/users/Androxys/pseuds/Androxys) “I do too, which is partially why I was so motivated to look into it. I'm looking at the relationship between "traditional literature" (whatever that means, but let's say "Pride and Prejudice") versus "professional transformative work," which I see as something like Neil Gaiman's "A Study in Emerald" or Seth Grahame-Smith's "Pride and Prejudice and Zombies" versus "amateur transformative work," which I see as being the vast majority of work hosted here--work done by someone out of love, largely without the motivations of the publication industry. Within that relationship, I'm especially interested in who has access to each tier, what those writers are like, and what sort of things are being written.”
2. User [Ellana42](https://archiveofourown.org/users/Ellana42/pseuds/Ellana42) “You just made my semester's datascience project muuuuch more fun :)”
3. User [Ahsoka323](https://archiveofourown.org/users/Ahsoka323/pseuds/Ahsoka323) “I’m working on my Bachelor’s in Data Analytics right now, and I’m so excited to be able to use this data in my Capstone project! I’ll let y’all know as soon as I get to it!”
4. User [GhostwithShotgun](https://archiveofourown.org/users/GhostwithShotgun/pseuds/GhostwithShotgun) “Amazing, just in time for my dissertation!”
5. User [kennagirl](https://archiveofourown.org/users/kennagirl/pseuds/kennagirl) “I'm a PhD student, and this data dump has become a central feature of my final project for both my empirical research class and my Cthuluscene class. It will also probably be a tent pole in my dissertation. These two spreadsheets are about to be my life for the next three years.”
6. User [Anderazu](https://archiveofourown.org/users/Anderazu/pseuds/Anderazu) “Neat, I was looking for something like this. I started prodding at the files myself, because this makes AMAZING procrastination for other things I was supposed to do this weekend. That's at <https://github.com/anderazu/ao3-stats/> in case it's useful for anyone (so far, just some exploratory frequency plots on the works).”
7. User [verity](https://archiveofourown.org/users/verity/pseuds/verity) “Hello! I recently did a little exploration of the characteristics of AO3 works with 75+ tags using this data set. Thanks for making it available!
   * [viz on tableau public](https://public.tableau.com/app/profile/ladyofthelog/viz/AO3TagsperWork2008-2020/Overview),
   * [write-up on github](https://github.com/ladyofthelog/ao3-data-dump).”
8. User Account Deleted “I've started a GitHub Repo where I'll be posting my R code on how I am cleaning / transforming the data. <https://github.com/mousemode/AO3_Data_Dump>”
9. User [fedormyskin](https://archiveofourown.org/users/fedormyskin/pseuds/fedormyskin) “I've already used similar data but this backend information will be really precious for my work <https://www.researchgate.net/publication/344495051_Cultural_accumulation_and_improvement_in_online_fan_fiction>”

**References**

### Archive of Our Own (2020, May 09). [AO3 Statistics 2020: A Look Behind the Scenes](https://archiveofourown.org/admin_posts/15931)

<https://archiveofourown.org/admin_posts/15931>

Archive of Our Own (2021, March 21). Selective data dump for fan statisticians.

<https://archiveofourown.org/admin_posts/18804>

Archive of Our Own (2022, January 27). Archive FAQ: Tags. Accessed January 27, 2023 at

https://archiveofourown.org/faq/tags?language\_id=en#wrangling

Castello, J. (2022, February 9). The *Lord of the Rings* Fanfic That’s 10 Times Longer Than the

Original. Accessed January 25, 2023,

<https://slate.com/culture/2022/02/lord-of-the-rings-longest-fanfiction-tolkien.html>

Johnson, S.F. (2014) "Fan Fiction Metadata Creation and Utilization within Fan Fiction

Archives: Three Primary Models." *Transformative Works and Cultures,* no. 17.

Noppe, N., Price, L., Chiu, K., Miller, J. N., Wang, E. N., Vaswani, S. M., Merry, S. K., Pollock,

D. E., Black, S. R., Hartwell, R., Jacobs, N., Thomas, P. A., Emmanouloudis, A., Hellman, E., & Spitz, A. (2022). What if academic publishing worked like fan publishing? Imagining the Fantasy Research Archive of Our Own. Transformative Works and Cultures, 37. <https://doi.org/10.3983/twc.2022.2253>

Price, L., & Robinson, L. (2021). Tag analysis as a tool for investigating information behaviour:

comparing fan-tagging on Tumblr, Archive of Our Own and Etsy. Journal of Documentation, 77(2), 320–358. <https://doi.org/10.1108/JD-05-2020-0089>