Where No One Has Gone Before: A Meta-Dataset of the World's Largest Fanfiction Repository

Kodlee Yin, Cecilia Aragon, Sarah Evans, Katie Davis
University of Washington, Seattle, USA
kodlee@fru1t.me, {aragon,sarahe,kdavis78}@uw.edu

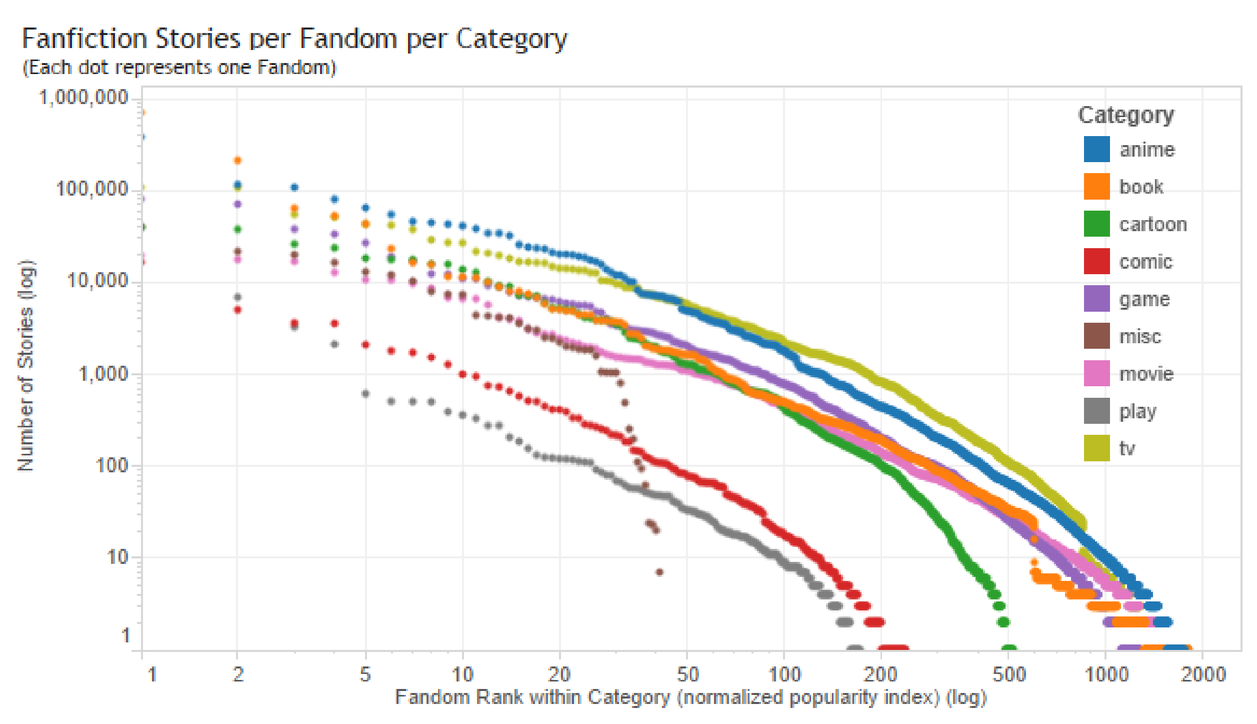

Figure 1. An overview of fanfiction story metadata by fandom/category. The category Books contains the most popular fandom (Harry Potter) by number of stories, but the category Anime is much deeper, with multiple fandoms generating more stories.

ABSTRACT

With its roots dating to popular television shows of the 1960s such as Star Trek, fanfiction has blossomed into an extremely widespread form of creative expression. The transition from printed zines to online fanfiction repositories has facilitated this growth in popularity, with millions of fans writing stories and adding daily to sites such as Archive Of Our Own, Fanfiction.net, FIMfiction.net, and many others. Enthusiasts are sharing their writing, reading stories written by others, and helping each other to grow as writers. Yet, this domain is often undervalued by society and understudied by researchers. To facilitate the study of this large but often marginalized community, we present a fully anonymized data release (via differential privacy) of the metadata from a large fanfiction site (to protect author privacy, story, profile, and review text is excluded, and only metadata is provided). We use visual analytics techniques to draw several intriguing insights from the data and show the potential for future research. We hope other researchers can use this data to explore further questions related to online fanfiction communities.

Author Keywords
Fanfiction; online communities; youth.

ACM Classification Keywords
H.5.3 Group and organization interfaces: Web-based interaction.

CHI 2017, May 6–11, 2017, Denver, CO, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-4655-9/17/05...$15.00.
http://dx.doi.org/10.1145/3025453.3025720

INTRODUCTION

Writing, sharing, and discussing fanfiction is a hugely popular online activity, and the site we call Fanfics.com (anonymized to comply with terms of service) is a vast online fanfiction repository. As of June 5, 2016, there were 6,402,752 stories, 166,739,153 story reviews, and 1,455,952 authors on the site. To put this last figure in context, there were 113,000 active editors on English Wikipedia on the same date. Yet the number of ACM papers published on Wikipedia far exceeds the number of papers studying fanfiction: 1,581 Wikipedia papers compared to a handful of papers that deal with fanfiction. We do not claim that fanfiction has had the impact or importance of Wikipedia; nevertheless, the phenomenon of fanfiction is clearly an understudied online context relative to its popularity.

It is not simply its popularity that makes fanfiction worthy of investigation; it is what people are doing on these sites—and how they are benefiting—that merits empirical inquiry. Authors (often young people) are publishing stories that are several hundred thousand words in length, sometimes exceeding the length of the original works on which they were based. Readers are offering encouragement and constructive feedback on stories, which authors use to improve their writing [5,9]. All of this happens in a positive, supportive community atmosphere that stands in stark contrast to the negativity and even hate speech that is found on so many online sites (e.g., Reddit, 4chan, comments posted on news sites).

Previous work has documented the benefits that people gain through their participation in online fanfiction communities. Authors experience mentorship from the community, grow as writers, gain recognition for their work, and form meaningful connections with other fans [2-5,9-18]. Most existing research on fanfiction has consisted of relatively small-scale, ethnographic investigations. These studies offer valuable insight into users' experiences and interactions in online fanfiction communities. However, they cannot provide a comprehensive view of the full scope of activity on a fanfiction repository. The current study provides such a view, constituting the first fully anonymized data release of a fanfiction site's complete metadata, consisting of six million stories across ten thousand fandoms. Previous work that undertook quantitative analysis only sampled the data [7,19,21]. We utilize differential privacy [8] for the data release so that researchers can explore questions related to online fanfiction communities without violating the privacy of the authors. A secondary contribution of this work is the demonstration of a visual analytics approach to draw insight from this vast corpus of human-generated data.

PREVIOUS WORK AND BACKGROUND

History of Fanfiction

Fanfiction is a type of transformative work in which fans of a variety of different media—including television shows, movies, comic books, anime, and video games—write stories that are based on the original media but expand upon or alter it. Fanfiction authors may consider plot trajectories different from the original story, explore the untold background stories of characters, or place the characters in an "alternate universe" that is completely different from the original setting. Fanfiction stories can vary in length from "one shots" of less than 1000 words to novel-length stories.

Fanfiction developed within the context of popular television shows of the 1960s, such as Star Trek and The Twilight Zone [11]. Yet, one might go considerably further back in history and regard the works of Shakespeare and Virgil as early examples of fanfiction. Well-known works of contemporary fanfiction include *Wicked: The Life and Times of the Wicked Witch of the West* (based on Baum's *The Wonderful Wizard of Oz*), *Pride and Prejudice and Zombies* (based on Austen's *Pride and Prejudice*), and *The Wind Done Gone* (based on Mitchell's *Gone with the Wind*). For every popular example of fanfiction, however, there are thousands more fanfiction stories that are unknown to the general public.

Research on Fanfiction

Research on fanfiction began nearly 25 years ago with Jenkins' seminal investigation of fan culture [11]. This work was pivotal in countering negative stereotypes of fans and fan culture by describing fan communities as rich sites of participatory culture. Jenkins challenged the common view that fanfiction was unoriginal, derivative work, instead positioning fanfiction stories as highly creative acts of manipulating and extending source material in unexpected, boundary-pushing ways.

Since Jenkins' foundational work, research on fanfiction has addressed issues of legality [20,22,23], identity [1,3,14], sexuality [11,24], and educational value [4,6,9,12,15-18]. For example, Black [2,3,4] explored English Language Learners' (ELL) participation in online fanfiction communities, focusing in particular on language acquisition and identity formation. More recently, Campbell et al. [5] conducted an in-depth nine-month ethnographic investigation of Fanfiction.net and FIMFiction.net, which included participant observations and interviews with authors. In a follow-up study, they conducted a thematic analysis of 4,500 reader reviews on Fanfiction.net [9].

Selecting ten original canonical works from Fanfiction.net, Milli and Bamman [19] employed natural language processing (NLP) methods to investigate differences in the characters emphasized and the treatment of gender in fanfiction stories compared to the original works. They found that secondary characters tended to be mentioned more frequently than primary characters in fanfiction stories. In addition, they found a small but significant increase in the proportion of female character mentions in fanfiction stories compared to the original works. This investigation represents the first to apply computational methods to the study of fanfiction data, illustrating the potential to derive empirical insights from such an approach. However, the data used by Milli and Bamma

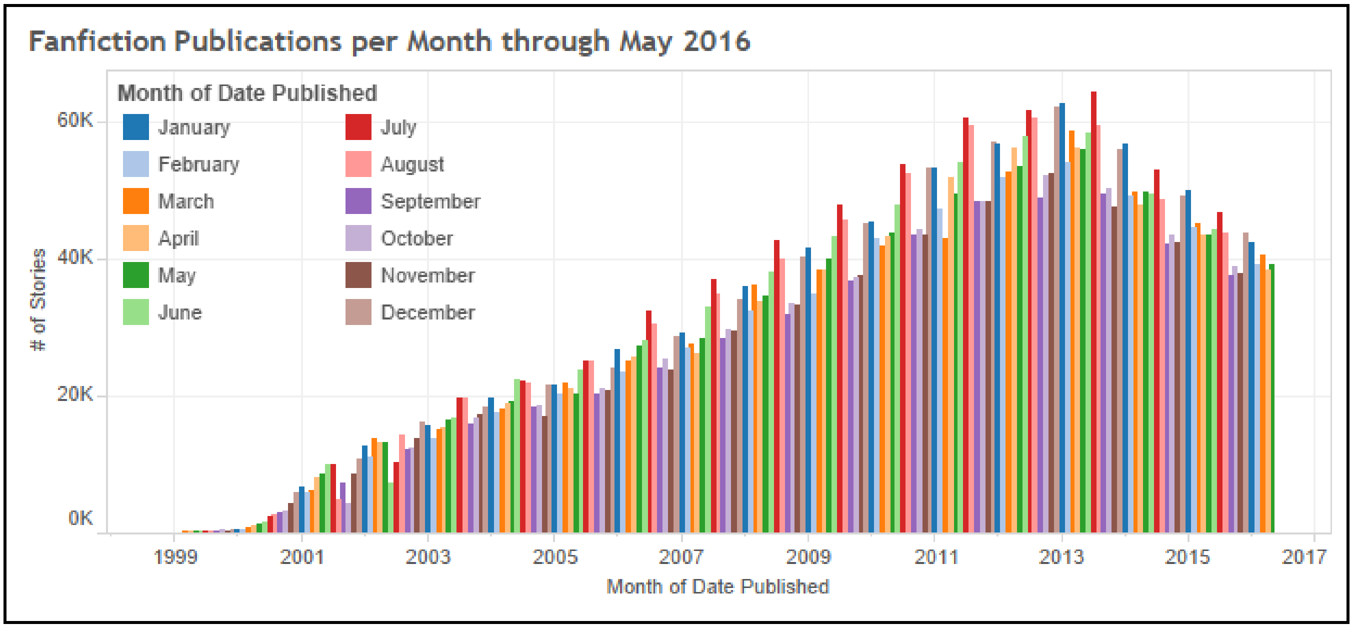

Figure 1. Story publication by month. Note the uptick in posting during northern hemisphere summer months.

constituted a fairly small proportion of the total corpus of fanfiction published on the site. The authors called for more quantitative analyses of fanfiction texts, and we hope that by making a large set of story metadata publically available and demonstrating its richness, we will be encouraging this type of research.

METHODS

Fanfics.com has various views for its website, none of which provide a complete snapshot. The site is organized hierarchically; the top-level subdivision is into nine *categories*: Anime/Manga, Books, Cartoons, Comics, Games, Misc, Movies, Plays/Musicals, and TV Shows. Each category is subdivided into *fandoms* (e.g. in the *Books* category, *Harry Potter* and *Twilight* are the two most popular fandoms). Finally, within each fandom, paginated lists of 25 stories per page, which we call *fandom pages*, provide the title, synopsis, author, content rating, language, and genre(s) (e.g., Romance, Drama) of each story, alongside statistics about each story such as word count, number of chapters, reviews, number of favorites and followers. Optional data includes the last time the story was updated, and characters from the fandom used in the story. Collectively, we define this information as the Fanfics.com *metadata*. It does not contain the story texts themselves, but is a rich and dynamic repository of information about the millions of writers and readers of fanfiction on this site.

We collected the data via a combination of Apache HttpComponents and jsoup. The list of fandoms was stored in a MySQL relational database. We processed the paginated fandom pages and stored the raw HTML results as-is for later processing. In total, we scraped and processed 10,339 fandoms on Sunday November 20, 2016. The data was stored in a normalized MySQL database with appropriate tables such as "genre," "story," "story_genre." Variables were aggregated and Laplacian noise added to ensure differential privacy even among outliers.

FINDINGS

The dataset is accessible at **http://research.fru1t.me**. It contains metadata describing 6,807,100 stories across 1,516,335 authors in 44 languages. Each row contains an anonymized user ID, story ID, six quantitative, and six categorical variables as documented on the release site. The stories span 10,294 fandoms and use 46,337 unique characters 9,308,807 times. The 20 genres were applied on 6,159,491 stories, leaving 647,609 uncategorized.

The largest categories were Anime/Manga, TV Shows, and Books with 1,905,055, 1,553,815, and 1,442,290 stories each, respectively. The top genres were Romance, Humor, and Drama with 3,423,862, 1,296,042, and 1,147,377 stories each, respectively. Finally, the top fandoms were *Harry Potter*, *Naruto*, and *Twilight* with 713,814, 387,218, and 212,929 stories, respectively.

We have provided graphs and Tableau visualizations of a few of our most interesting findings. Figure 1 illustrates the differing "shapes" of fandom popularity within categories. Although the category Books contains the single most popular fandom (*Harry Potter*) by number of stories, the category Anime is much deeper, with multiple fandoms generating more stories overall.

We also observe that the hypothesis, based on our previous ethnographic investigations, that a significant portion of the authors on this site are English-speaking students, appears to be confirmed by the uptick in posting during the northern hemisphere summer months, with a brief blip in late December (Figure 2). Note the steady growth until 2013; we have no explanation for the drop-off after that year, and further research is encouraged.

Category	Stories	Authors	Avg Words	Median Words	Words StDev.	Avg. Chptrs	Median Chptrs	Chptrs StDev.	Avg. Reviews	Median Reviews	Reviews StDev.
anime	1,876,647	500,902	8,838	2,379	24,639	3.79	1	7.29	24.3	7	93.4
book	1,422,285	488,670	10,009	2,287	27,586	4.56	1	8.32	39.7	7	217.5
cartoon	453,925	138,790	8,885	2,412	25,003	4.23	1	7.28	20.4	7	59.8
comic	52,267	24,924	7,117	1,935	22,021	3.39	1	6.59	12.9	4	52.2
game	634,519	231,497	9,814	2,372	29,887	4.11	1	8.36	14.7	4	91.1
misc	207,563	101,161	7,544	1,929	37,524	3.83	1	8.26	12.0	3	48.1
movie	291,075	117,251	9,579	2,528	24,178	4.70	1	8.31	23.9	7	77.1
play	60,227	17,958	6,319	2,371	12,908	2.79	1	4.28	41.4	11	132.7
tv	1,404,244	299,347	8,867	2,296	24,040	4.18	1	8.56	23.8	7	79.7

Table 1. Per category information for the fanfiction database. Note that Anime is the most popular overall category, but each Anime fanfiction story receives fewer reviews on average than Book fanfiction.

	Adventure	Angst	Comfort/Hurt	Crime	Drama	Family	Fantasy	Friendship	Horror	Humor	Mystery	Parody	Poetry	Romance	Sci-Fi	Spiritual	Supernatural	Suspense	Tragedy	Western
Adventure	0	1061	324	55	3663	1341	2871	2437	205	3836	2008	131	42	13140	182	60	673	707	273	4
Angst	1061	0	3496	48	8741	2072	222	1312	612	994	367	123	1223	25148	11	186	247	356	3876	2
Comfort/Hurt	324	3496	0	25	1656	5231	154	3727	72	395	113	22	219	11924	1	87	82	85	1292	2
Crime	55	48	25	0	84	34	17	29	24	49	49	6	7	100	1	1	8	25	45	0
Drama	3663	8741	1656	84	0	2121	749	1449	435	3409	973	148	427	31054	21	141	354	799	1896	2
Family	1341	2072	5231	34	2121	0	338	2783	64	3219	175	32	104	5823	16	63	92	99	954	5
Fantasy	2871	222	154	17	749	338	0	593	52	896	301	68	66	2826	97	33	281	152	122	0
Friendship	2437	1312	3727	29	1449	2783	593	0	32	3973	192	37	78	12769	15	64	73	82	431	0
Horror	205	612	72	24	435	64	52	32	0	341	167	54	59	538	21	18	143	148	271	1
Humor	3836	994	395	49	3409	3219	896	3973	341	0	603	7375	525	46213	52	50	221	209	278	17
Mystery	2008	367	113	49	973	175	301	192	167	603	0	31	33	2511	12	21	145	305	117	2
Parody	131	123	22	6	148	32	68	37	54	7375	31	0	110	607	2	7	18	21	50	7
Poetry	42	1223	219	7	427	104	66	78	59	525	33	110	0	1971	1	42	22	15	392	1
Romance	13140	25148	11924	100	31054	5823	2826	12769	538	46213	2511	607	1971	0	80	251	765	1018	4943	15
Sci-Fi	182	11	1	1	21	16	97	15	21	52	12	2	1	80	0	5	16	11	6	0
Spiritual	60	186	87	1	141	63	33	64	18	50	21	7	42	251	5	0	53	8	127	2
Supernatural	673	247	82	8	354	92	281	73	143	221	145	18	22	765	16	53	0	84	51	1
Suspense	707	356	85	25	799	99	152	82	148	209	305	21	15	1018	11	8	84	0	129	0
Tragedy	273	3876	1292	45	1896	954	122	431	271	278	117	50	392	4943	6	127	51	129	0	0
Western	4	2	2	0	2	5	0	0	1	17	2	7	1	15	0	2	1	0	0	0

Table 2. Combinations of genres in the Harry Potter fandom. Humor/Romance is the most popular combination.

A simple table of category-aggregated metrics (Table 1) reveals interesting facts at a glance: Anime is the most popular category, but stories in Books and Plays receive more reviews per story on average, indicating, as has been shown in previous research [5, 9], more engagement and support of authors in those categories. Stories are significantly longer in the Books category than they are in any other category.

We extracted multiple visualizations from the data, only a small fraction of which are reproduced here due to space constraints. We found a classic power-law distribution in the word counts of stories with a peak under a thousand words. A matched-genre heat map table for the *Harry Potter* fandom (Table 2) reveals both understandable and intriguing correlations. It is not surprising that Romance contains more stories than any other genre, but it is intriguing that Western has the least. It is also interesting that Humor/Romance was the most popular combination of genres, comprising 7% of all *Harry Potter* stories out of 400 possible combinations.

CONCLUSION

With this work, we hope to make generally available a fully anonymized, rich dataset about a highly popular but understudied phenomenon. Even a cursory glance through the data uncovers fascinating details about differences among fandoms, varying proportions of stories between categories, how genre types vary by language, and more. We hope this dataset will pique researchers' curiosity and spur further study of this topic.

ACKNOWLEDGMENTS

We thank the reviewers for their helpful comments.

REFERENCES

1. Ayşegül Kuglin Altintaş. 2013. A New Hermione: Re-Creations of the Female Harry Potter Protagonist in Fan Fiction. *Z Anglist Am* 61, 2: 155–173.
2. Rebecca W. Black. 2006. Language, culture, and identity in online fanfiction. *E–Learning* 3, 2, 170-184.
3. Rebecca W. Black. 2007. Digital design: English language learners and reader reviews in online fiction. In *A new literacies sampler*, Michele Knobel and Colin Lankshear (eds.). Peter Lang, New York, NY, 115-136.
4. Rebecca W. Black. 2008. *Adolescents and Online Fan Fiction*. Peter Lang, New York, NY.
5. Julie Campbell, Cecilia Aragon, Katie Davis, Sarah Evans, Abigail Evans, and David Randall. 2016. Thousands of positive reviews: distributed mentoring in online fan communities. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW '16). http://dx.doi.org/10.1145/2818048.2819934
6. Kelly Chandler-Olcott and Donna Mahar. 2003. Adolescents' *anime*-inspired "fanfictions": An exploration of multiliteracies. *J Adolesc Adult Lit.* 46, 7, 556-566.
7. Abigail De Kosnik, Laurent El Ghaoui, Vera Cuntz-Leng, Andrew Godbehere, Andrea Horbinski, Adam Hutz, Renee Pastel, and Vu Pham. 2015. Watching, creating, and archiving: Observations on the quantity and temporality of fannish productivity in online fan fiction archives. *Convergence: The International Journal of Research into New Media Technologies,* 21, 1:145-164.
8. Cynthia Dwork. A firm foundation for private data analysis, *Communications of the ACM*, vol. 54, 2011.
9. Sarah Evans, Katie Davis, Abigail Evans, Julie Campbell, David Randall, Kodlee Yin, and Cecilia Aragon. 2017. More than peer production: fanfiction communities as sites of distributed mentoring. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW '17). http://dx.doi.org/10.1145/2998181.2998342
10. Casey Fiesler, Shannon Morrison, and Amy S. Bruckman. 2016. An archive of their own: a case study of feminist HCI and values in design. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16). https://doi.org/10.1145/2858036.2858409
11. Henry Jenkins. 1992. *Textual Poachers: Television Fans & Participatory Culture.* Routledge, New York, NY.
12. Henry Jenkins. 2006. *Convergence Culture: Where Old and New Media Collide*. New York University Press.
13. Shannon Fay Johnson. 2014. Fan fiction metadata creation and utilization within fan fiction archives: Three primary models. *Transform Work Cultures* 17.
14. Soomin Jwa. 2012. Modeling L2 writer voice: Discoursal positioning in fanfiction writing. *Computers and Composition* 29, 4: 323-340.
15. Jayne Lammers. 2013. Fangirls as teachers: Examining pedagogic discourse in an online fan site. *Learning, Media and Technology* 38,4: 368-386.
16. Jayne Lammers. 2016. "The Hangout was serious business": Leveraging participation in an online space to design Sims fanfiction. *Res Teach Engl.* 50,4: 309-332.
17. Jayne C. Lammers and Valerie L. Marsh. 2015. Going public: An adolescent's networked writing on fanfiction.net. *J Adolesc Adult Lit*. 59, 3: 277-285.
18. Kerri Mathew and Devon Adams. 2009. I love your book, but I love my version more: Fanfiction in the English language arts classroom. *ALAN Review* 36, 3: 35-41.
19. Smitha Milli and David Bamman. 2016. Beyond canonical texts: a computational analysis of fanfiction. In *Proceedings of the Empirical Methods on Natural Language Processing Conference* (EMNLP '16).
20. Mollie E Nolan. 2006. Search for original expression: Fan fiction and the fair use defense. *SIU Law J* 30, 3: 533-571.
21. Charles Sendlor. 2011. Fan Fiction Demographics in 2010: Age, Sex, Country. Retrieved September 17, 2016 from http://ffnresearch.blogspot.com/2011/03/fan-fiction-demographics-in-2010-age.html
22. Leanne Stendell. 2005. Fanfic and fan fact: How current copyright law ignores the reality of copyright owner and consumer interests in fan fiction. *SMU Law Rev* 58: 1551.
23. Rachel L Stroude. 2010. Complimentary creation: Protecting fan fiction as fair use. *Marquette Intellect Prop Law Rev* 14, 1: 191.
24. Catherine Tosenberger. 2008. Homosexuality at the online Hogwarts: Harry Potter slash fanfiction. *Child Lit* 36, 1: 185–207.