Twilight Fansites: Sarah Green, Rachel Mauro, Connor Wynn, Kristen Taynor

Collection Description

We chose to archive fansites and other online information pertaining to the "Twilight" franchise. "Twilight" represents a modern-day phenomenon with movie studios picking up popular, young adult fantasy book series to adapt into multi-part films. This is due, as our report shows, to the enthusiastic response by fans who will not only spend their money in book stores and movie theaters but who create a voluminous amount of fan materials, primarily stored and shared on the Internet. "Twilight," therefore, brings together many humanities disciplines, from literature to film studies to cultural studies as a start. The cultural impact alone of the franchise that these sites demonstrate makes this collection potentially of interest to future scholars and thereby worthy of preservation. Specifically, we also chose "Twilight" over similar franchises due to the final adaptation, *Breaking Dawn—Part Two*, hitting theaters in the middle of our web archiving assignment.

Collection Seeds

- Twilight Guide (http://twilightguide.com/tg/): fansite offering shopping guide, movie countdowns, pictures & graphics, quotes, quizzes, polls and a blog.
- Twilight Lexicon (http://www.twilightlexicon.com/): fan-run site serving as a reference to the Twilight books and movies, including chapter summaries, Twilight-saga mythology, email correspondence with Stephenie Meyer, and news about those involved in making the Twilight movies.
- Twilight Saga Wiki (http://twilightsaga.wikia.com/wiki/Twilight_Saga_Wiki): fan-run Wiki using Wikia technology to run a franchise-specific reference resource that can be edited by the community.

- Twilighters—Twilight Fansite Evolved (http://twilighters.org/): fansite featuring a blog, news, forums, and detailed information on the books and especially the movies.
- "Twilight" fanfiction on Fanfiction.net (http://www.fanfiction.net/book/Twilight/): sub-page of a multi-genre repository where fans can store and tag their "Twilight" fanfiction.
- Team Twilight (http://team-twilight.com/): fansite featuring blog, videos, polls, contest and chat.
- The Twilight Saga movie website (http://www.breakingdawn-themovie.com/): official movie website for all five films, also including links to other merchandise and embedded fan reactions from social media sites.
- The Twilight Saga forums (http://thetwilightsaga.com/forum/): fan forums organized by category, including specific characters, novels, the movies as a whole, and off-topic offerings.

Scoping

Our group started with the home URLs of eight seeds, then added three more seeds later in the project when we thought we needed to have exactly ten, which was an error on our part. We narrowed our focus back to the eight original seeds when we realized eight was enough. With each crawl of thetwilightsaga.com, we had large quantities of queued files that did not decrease over time, which meant that each crawl was extended indefinitely until it either timed out or was stopped by us. After looking at the files archived from the first production crawl, we decided to block the extension /activity/log/ and /main/authorization/ for thetwilightsaga.com because

1) URLs with those strings tended to be blocked by robots.txt, and 2) the pages appeared to need a username and password to access, rendering them inaccessible regardless of robots.txt code.

We noticed as well from the first production crawl report that the websites crawled with the highest number of out-of-scope and lowest number of in-scope pages were located at YouTube, Facebook, and Twitter URLs. After taking a peek at the "in scope" Twitter and Facebook pages, we decided that preserving them was unnecessary because many of the pages were for profiles set to private, were for profiles that only occasionally referred to Twilight, or were "share" links. So, in the hope of shortening crawl time, we set rules to block these networking sites, as well as some related YouTube sites and files like youtu.be and ytimg.com. Blocking Facebook and YouTube and its related files may not have been the best decision, however, as our late QA reports (the QA function seemed to be unavailable for at least a week) showed that many embedded videos and comments did not render in our archived sites.

Even with our adjustments, the crawls still never completed on their own. This meant that we had to stop them ourselves or let them time out. In a last-ditch effort, we decided to try limiting the scope of thetwilightsaga.com (our most troublesome site) to just the site's discussion forum. This meant changing thetwilightsaga.com seed URL to thetwilightsaga.com/forum/. The software initially ran into an HTTP error 500 for that URL, but a patch crawl appeared to fix that issue and we were able to capture the forum pages.

File Formats Preserved

For the eight currently active seeds, the majority of the files we captured were overwhelmingly HTML text pages that ranged within the mid-to-upper 100,000s. For other file formats, numbers dropped dramatically for the fourth and final crawl, probably because of the dramatic scoping restrictions we placed on thetwilightsaga.com, and possibly because the scoping rules we placed slowed the crawl engine. A large number of JPEG image files were crawled each time, ranging between 100,000 and 150,000 for the first two crawls (Fig. 1 and 2)

but dropping to about 60,000 by the last crawl (Fig. 3). About 70,000 “progressive” JPEG (PJPG) files were collected from the first two crawls, but numbered seven in the fourth crawl. XML text files numbered 26,000 in the first crawl all the way up to 100,000 in the fourth crawl. Atom+XML application files counted around 30,000 for the first two crawls, but dropped to 16 in the last crawl. RSS+XML application files numbered at around 14,000 files (Fig. 1), then 920 (Fig. 2), then insignificant amounts in the fourth crawl (Fig. 3).

The video file numbers were very low in general. The most numerous file format was the mp4, which numbered around three on average, followed by .avi and .x-flv file formats. The numbers of .PDF files were low as well, ranging between 0 and 30, depending on the crawl.

Troubleshooting

The sheer size of our collection led to an obvious but perhaps overlooked rendering issue—not knowing what had or had not rendered properly. At the time of this writing, the class in total has archived just over nine million pages for seven unique collections; of those pages, nearly four million are from our collection alone. Even a spot-check of pages would only put a small dent in that number, which means we simply do not know everything that did or did not render properly. We have run patch crawls for all production crawls containing our eight active seeds, now that the QA reports have finished, in order to make a best effort to fix items the crawler knows have rendering issues.

Another rendering issue stems from the fact that we blocked as much of the YouTube, Facebook, and similar networking content as we could isolate, as already discussed. This means that the YouTube videos and Facebook content either has not rendered properly or shows up but does not work on the pages when viewed in the Wayback Machine. Similarly, the other scoping modifiers we placed on seeds will cause those items and pages to not render properly.

Finally, not all of our seeds were fully crawled because none of our production crawls finished on their own. Two of the production crawls stopped due to a self-imposed Archive-It time limit of three days; our group stopped the other two production crawls because they had archived more than one million documents each. Three of our seeds, team-twilight.com, thetwilightsaga.com, and twilightguide.com, consistently had a large number of documents still in the queue when the crawl either timed out or was stopped and will be incomplete in their archived versions as will any other pages still in the queue when the crawl stopped. Unlike the QA reports that allow us to troubleshoot known rendering issues and run patch crawls, no fix exists for the crawls timing out, other than reducing the number of seeds or creating even more crawl modifiers than we already had in place.

Lessons Learned

The major lessons we learned, due to the size of our collection, relate to prioritizing and compromising. If we were to redo this assignment in the future we might consider running a crawl of each of the larger seeds separately. This might mean that each larger seed would only get crawled once, but taking a large-picture view, it might be a higher priority to get one complete snapshot of a website rather than several incomplete ones. Another compromise was to limit content in the hopes of actually finishing a production crawl. This led to a trade-off of having rendering issues and missing content that while potentially valuable was causing the crawls to time out and leave many more valuable documents in the queue (Fig. 4).

The rendering issues and scoping decisions we had to troubleshoot make it clear that the web truly is a mess. Lori Donovan's presentation slides show an image of a massive web of cords and wires when describing the web as a mess (slide 25) and this point was driven home in the number of pages that our collection attempted to archive. On its surface, each seed looks

harmless, but they are each unique creations with a surprisingly complex mass of pages the further you mine into each. We mostly archived fansites that are designed to be organic and do not have a finite purpose, which means they hide a voluminous amount of forum, chat, and other user-input pages; images, videos, and other multimedia; and layer upon layer of movie- and book-related pages. But even with the web being a mess and us not being able to completely render some seeds, a humanist sometime in the future may be appreciative for the content we were able to preserve.

Appendix

Figure 1: Screen cap of most numerous file formats of first crawl, which ran from November 15-18, 2012.

File Type	URLs ↓	Data
text/html	381162	16.7 GB
image/jpeg	111937	2.4 GB
image/pjpeg	69060	949.7 MB
application/atom+xml	31687	319.0 MB
text/xml	26084	98.1 MB
image/gif	15432	2.9 GB
application/rss+xml	14014	132.1 MB
image/png	8998	635.4 MB
text/dns	8255	996.4 KB
application/x-shockwave-flash	6711	263.9 MB
image/x-png	4127	268.1 MB
text/plain	4017	142.4 MB
text/javascript	1660	34.6 MB
image/bmp	1444	334.5 MB
no-type	904	1.7 MB
text/css	684	12.1 MB
application/xml	538	738.1 KB
application/x-www-form-urlencoded	95	476.2 KB
application/octet-stream	90	37.4 MB
application/x-javascript	62	3.5 MB
image/GIF	41	385.8 KB
image/x-bmp	20	974.6 KB
binary/octet-stream	13	1.9 MB
application/javascript	11	181.6 KB

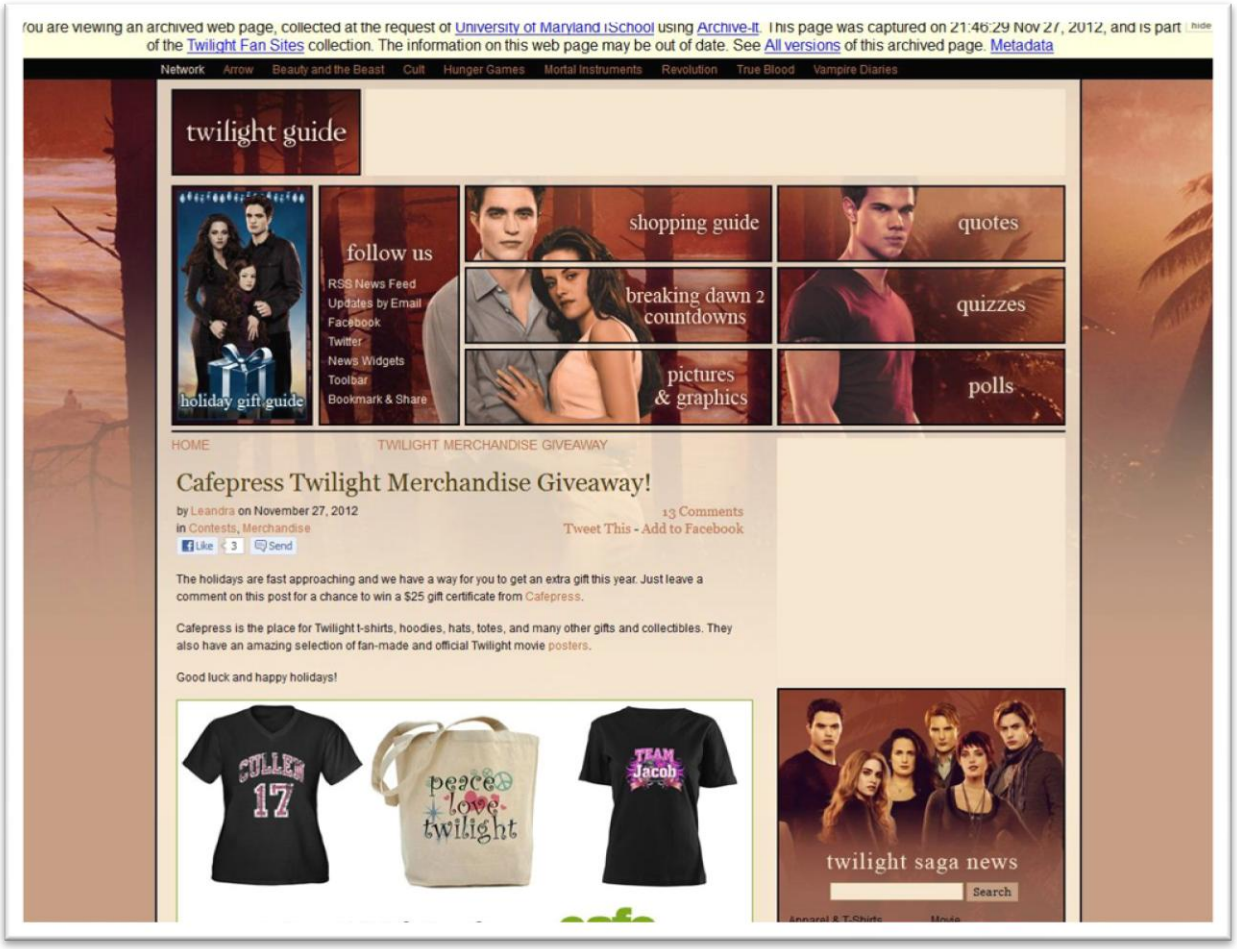
Figure 2: Screen cap of most numerous file formats of second crawl, which crawled only five of the eight seeds, and ran from November 18-21, 2012.

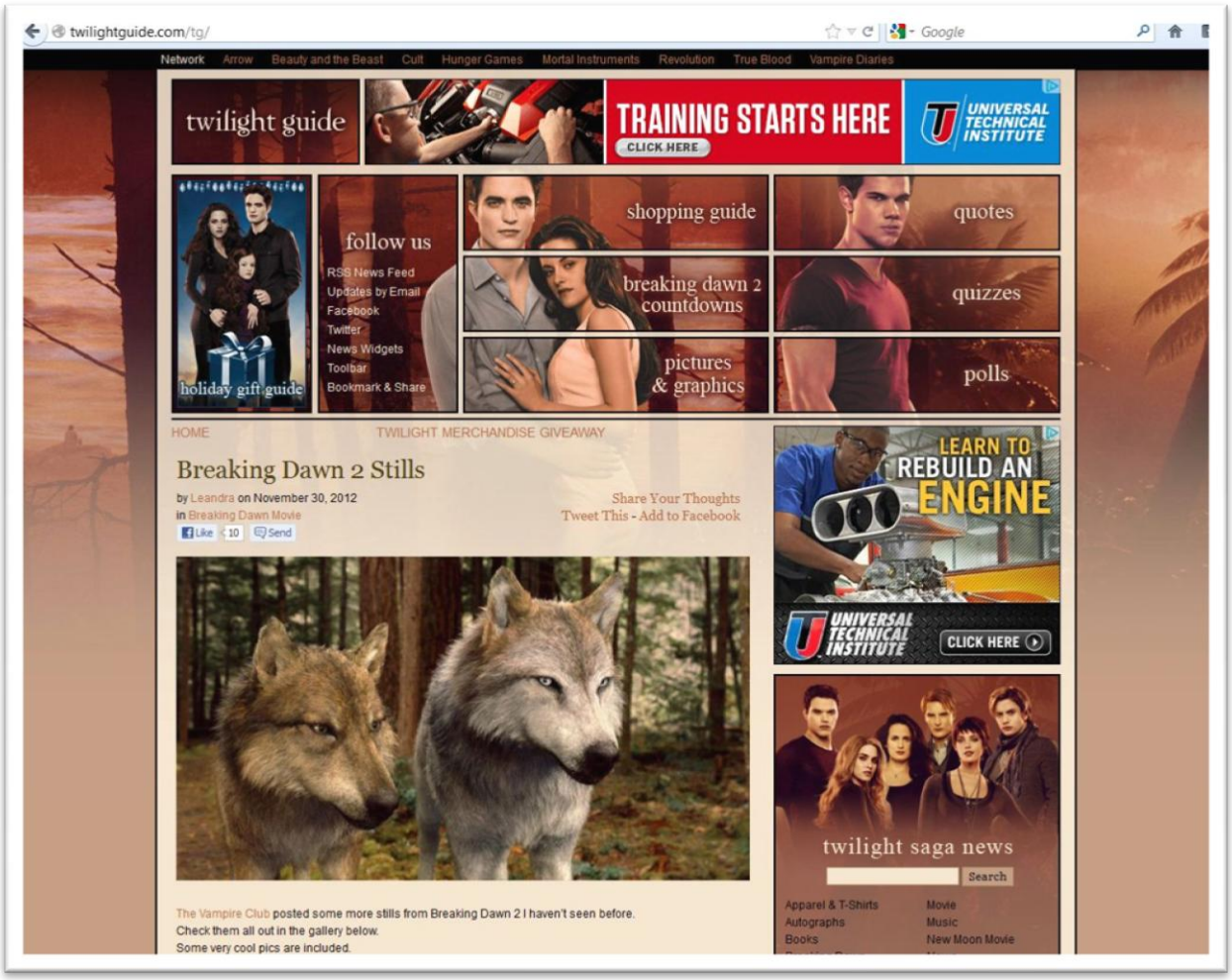
File Type	URLs ↓	Data
text/html	800177	22.8 GB
image/jpeg	141885	4.2 GB
text/xml	97343	473.4 MB
image/pjpeg	70931	971.1 MB
application/atom+xml	25965	261.4 MB
image/gif	16346	3.1 GB
text/javascript	13989	43.9 MB
image/png	12446	1.1 GB
text/dns	11239	1.3 MB
application/x-shockwave-flash	9231	267.4 MB
text/plain	5350	171.2 MB
image/x-png	4190	254.1 MB
image/bmp	1405	332.7 MB
no-type	929	1.4 MB
application/rss+xml	920	15.8 MB
text/css	619	11.0 MB
application/xml	602	753.5 KB
application/octet-stream	120	49.4 MB
application/x-www-form-urlencoded	111	558.5 KB
application/x-javascript	100	4.3 MB
image/jpg	36	1.4 MB
application/pdf	32	12.6 MB
image/x-bmp	28	1.8 MB

Figure 3: Screen cap of most numerous formats of fourth crawl, which ran from November 27-30, 2012, and included the eight original seeds from the first crawl.

File Type	URLs ↓	Data
text/html	556058	17.4 GB
text/xml	104140	490.2 MB
image/jpeg	61099	3.0 GB
text/javascript	14953	18.2 MB
image/gif	4917	478.3 MB
text/dns	4669	513.0 KB
application/x-shockwave-flash	4420	108.4 MB
image/png	4388	565.8 MB
text/plain	2408	115.5 MB
no-type	225	70.5 KB
application/xml	158	158.0 KB
application/x-javascript	71	1.8 MB
image/bmp	46	19.9 MB
text/css	44	801.5 KB
application/octet-stream	43	801.5 KB
application/pdf	25	8.3 MB
application/javascript	22	306.6 KB
application/vnd.javascript	20	565.2 KB
application/atom+xml	16	72.6 KB
image/x-icon	11	12.9 KB
application/xhtml+xml	8	179.7 KB
image/pjpeg	7	375.9 KB
video/mp4	5	32.8 MB
image/jpg	3	38.3 KB

Figure 4: Pair of screen caps showing twilightguide.com in archived and live form, respectively. The archived version, from the fourth and final crawl on November 27, does not contain advertisements, which is some of the content we decided to block for scoping purposes.




twilightguide.com/tg/
Google
Network
Arrow
Beauty and the Beast
Cult
Hunger Games
Mortal Instruments
Revolution
True Blood
Vampire Diaries
twilight guide
TRAINING STARTS HERE
CLICK HERE
UNIVERSAL TECHNICAL INSTITUTE
holiday gift guide
follow us
RSS News Feed
Updates by Email
Facebook
Twitter
News Widgets
Toolbar
Bookmark & Share
shopping guide
quotes
breaking dawn 2 countdowns
quizzes
pictures & graphics
polls
HOME
TWILIGHT MERCHANDISE GIVEAWAY
Breaking Dawn 2 Stills
by Leandra on November 30, 2012
in Breaking Dawn Movie
Like 10 Send
Share Your Thoughts
Tweet This - Add to Facebook
The Vampire Club posted some more stills from Breaking Dawn 2 I haven't seen before.
Check them all out in the gallery below.
Some very cool pics are included.
LEARN TO REBUILD AN ENGINE
UNIVERSAL TECHNICAL INSTITUTE
CLICK HERE
twilight saga news
Search
Apparel & T-Shirts
Autographs
Books
Movie
Music
New Moon Movie