The GOLEM Triple Store: A Graph-based Representation of Narrative and Fiction

Franziska Pannach[1,*], Xiaoyan Yang[1], Noa Visser Solissa[1], Ze Yu[1], Andreas van Cranenburgh[1], Michiel van der Ree[2] and Federico Pianzola[1]

[1]Center for Language and Cognition (CLCG), University of Groningen
[2]Center for Information Technology (CIT), University of Groningen

Abstract

In this paper, we present the GOLEM triple store, a massive triple store resource for fiction and narrative. This triple store is the first step towards a large-scale knowledge-graph for stories, as well as characters and events in narratives. At the moment, it contains more than 8 million stories collected from the Archive of Our Own (AO3) [1], providing scholars with a tool to derive unique insights into fan narratives and storytelling trends over time.

1. Introduction

In this article we introduce a new resource for the large scale study of fiction on the basis of metadata and "derived data" [2] – or "mesodata" [3] – that is, various textual features that allow to compare documents without accessing their full text. The idea is similar to that of the HathiTrust Extracted Features dataset [4], but the features encoded in the GOLEM ("Graphs and Ontologies for Literary Evolution Models") triple store are much richer, also referring to narrative and stylistic elements, and to reader response data (e.g. characters, relationships, topics, readability, sentiment of comments received by the story, etc.). Similar projects exist on a smaller scale for a selection of texts in English [5], Dutch [6] and German [7]. The creation of the GOLEM triple store has been inspired by such work but will operate on a completely different scale, which requires the automation of the extraction of textual features for millions of stories.

The core concept of the GOLEM infrastructure is that of "programmable corpora", i.e. "research-oriented corpora providing an API" [8], which allows to easily reapply scripts, notebooks, and pipelines of analysis to all texts in the corpora, inasmuch as they are encoded following the same principles and can be queried via the same API and SPARQL endpoint. Since the GOLEM focuses primarily on derived data, there is no need for a resource-intensive XML database of texts encoded in TEI[1] format, like that created by [8]. Only statements about the texts and their reception are stored in the database.

Semantic Methods for Events and Stories (SEMMES) Workshop, 2024
*Corresponding author.
f.a.pannach@rug.nl (F. Pannach); xiaoyan.yang@rug.nl (X. Yang); noa.visser@rug.nl (N. V. Solissa); z.yu@rug.nl (Z. Yu); a.w.van.cranenburgh@rug.nl (A. v. Cranenburgh); michiel.van.der.ree@rug.nl (M. v. d. Ree); f.pianzola@rug.nl (F. Pianzola)

[1]Text Encoding Initiative

The first batch of texts made available in the GOLEM triple store are gathered from the largest and most popular fanfiction platform, Archive of Our Own (AO3)[2]. The spread of digital and social media in the 21st century has reconfigured many literary dynamics, namely it has reduced the influence of literary institutions like literary critics, publishers, and schools, creating more spaces and occasions for niche and amateur fiction to become popular, thanks to reader-to-reader interactions. Fanfiction platforms and website for book reviews/discussion (e.g. Goodreads) are now some of the most thriving environments to study narrative, fiction, and reader response. In the fanfiction space, readers become writers, viewers become creators and recipients become participants of their favorite (fictional) universes. Since fanfiction writers publish their own works independently from publishing houses and editors, their creativity knows no bounds, and is not subjected to limitations or censorship. Writers easily cross from one fictional universe (fandom) into another, or place themselves (or the reader) as characters in their favorite stories. Fanfiction has become an integral part of transformative fan culture, a cultural phenomenon in its own right.

Up to 2022 (the cut-off point of our data collection), more than 8.7 million stories were published on AO3 in the English language alone. Additionally, there is a wide coverage of other languages, including Chinese, Italian, or Korean including many so-called low-resource languages, such as Bahasa Indonesian or isiZulu. Therefore, the fanfiction domain holds immense potential for the study of user-produced narratives, readers' response, semantic and narrative modeling approaches, and for the development of natural language processing (NLP) tools for low- or under-resourced languages. While certain individual features of fanfiction domain have been investigated, such as user-selected and user-provided tags [9], the narratives in their entirety are largely under-studied.

The GOLEM Project triple store is a large and easily accessible resource for querying AO3 data and more data sources will be added later on. We demonstrate the potential of the triple store in three case studies.

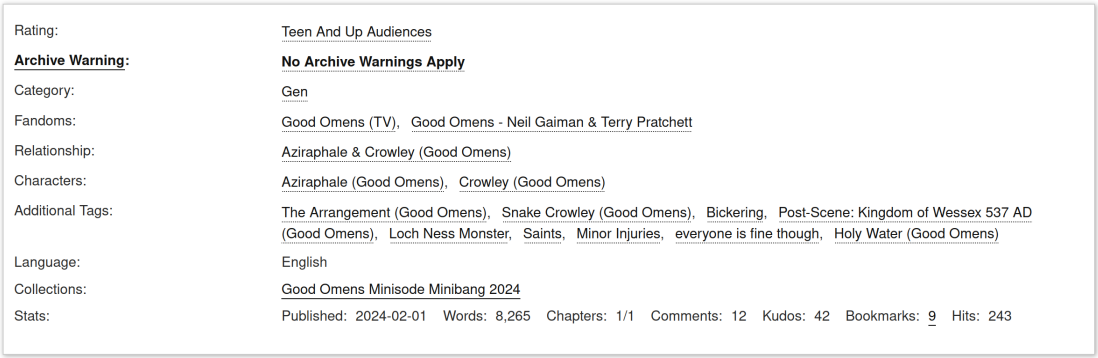
The article is structured as follows: Section 2 describes the data and its representation in the triple store. Section 3 presents some illustrative case studies. Section 4 contains the discussion, and Section 5 describes future work and planned extensions.

2. Data

Apart from the textual data provided by fanfiction writers and the common metadata such as author and title, AO3 provides a wide array of additional metadata, such as user-selected content tags, characters appearing in the story, as well as their relationships. Particularly popular are romantic (canon and non-canon) character pairings. Users can praise and react to each other's work by giving kudos or leaving comments.

Individual stories in the triple store are identified by their story ID. Each story has a number of associated metadata items, such as summary, word count, date published and more. As of date, all predicates in the triple store are using the golem prefix (*https://golemlab.eu/graph/*), derived in parts by properties from *CIDOC-CRM* [10], *Schema.org* [11], and *LRMoo* [12]. The triple store maintains the user-selected (upper case) and user-generated (lower case) tags via

[2]https://archiveofourown.org



Rating:	Teen And Up Audiences
Archive Warning:	**No Archive Warnings Apply**
Category:	Gen
Fandoms:	Good Omens (TV), Good Omens - Neil Gaiman & Terry Pratchett
Relationship:	Aziraphale & Crowley (Good Omens)
Characters:	Aziraphale (Good Omens), Crowley (Good Omens)
Additional Tags:	The Arrangement (Good Omens), Snake Crowley (Good Omens), Bickering, Post-Scene: Kingdom of Wessex 537 AD (Good Omens), Loch Ness Monster, Saints, Minor Injuries, everyone is fine though, Holy Water (Good Omens)
Language:	English
Collections:	Good Omens Minisode Minibang 2024
Stats:	Published: 2024-02-01 Words: 8,265 Chapters: 1/1 Comments: 12 Kudos: 42 Bookmarks: 9 Hits: 243

Figure 1: Example Metadata for AO3 data, Source: https://archiveofourown.org/

the predicate *keyword*. Notably, the GOLEM triples store does not provide access to the full text, which remains in solely accessible through the Archive of Our Own. However, text-based features with regard to events and character-features are planned to be incorporated in the knowledge graph, see Section 6. Table 1 explains the relevant data fields and gives examples where needed. Some values that were originally aggregated in lists, such as additional tags in Figure 1, are split into multiple triples, e.g. *golem:keyword* to facilitate explorability of the data (like querying specific keywords in fandoms).

For internal use, the data was first harvested from the archive.org archive[3] and stored in an internal Elasticsearch database with help of a custom ingest script[4]. This database has now been converted into triple store data and is available at: http://graph.golemlab.eu:8890/sparql via an institutional Virtuoso server[5]. Virtuoso was chosen because it scales well with growing knowledge graphs. Even holding multiple billions of triples on a single instance, single machine setup, Virtuoso still performs well, in a real-life setting.[6].

Up to the date of this publication, metadata for 8 million stories have been made available. The data in the GOLEM triple store contains the story related metadata in AO3 up to and including December 2022. With this choice we want to limit the stories that are potentially written with the aid of large language models, allowing for a more reliable investigation of *human* storytelling. While comparing human-generated stories with narratives produced by large language models could be an interesting area of research, it is not currently within the scope of the project. Extending the knowledge graph with more recent (human) user-produced stories from AO3 and other fanfiction platforms is planned for future updates.

[3] https://archive.org/details/AO3_final_location
[4] Available at https://github.com/GOLEM-lab/golem-ingest
[5] https://virtuoso.openlinksw.com/
[6] See UniProt https://www.w3.org/wiki/LargeTripleStores#OpenLink_Virtuoso_v7.2B_.2894.2B.2B_explicit.2C_uncounted_virtual.2Finferred.2C_in_1_instance_on_1_machine.29

Table 1
Triple Store Predicates

Predicate	Explanation	Example
golem:author	Username Author (anonymised)	
golem:characters	Characters appearing in the story	Molly Weasley
golem:collections	Title of the collection that a story is part of	Good Omens Minisode Minibang 2024
golem:contentWarning	Content warnings regarding level of violence/sexuality	Graphic Depictions Of Violence
golem:datePackaged	Date packaged for the project database	
golem:datePublished	Date published on AO3	
golem:dateModified	Date updated by the author	
golem:fandom	Fictional universe(s) of the story	Good Omens (TV Show)
golem:keyword	User-provided content keywords	Loch-Ness Monster
golem:language	Language in which the story is written	English, Italiano
golem:numberOfChapters	Number of chapters	
golem:numberOfComments	Number of comments	
golem:numberOfKudos	Number of user-approvals (similar to likes)	
golem:numberOfWords	Number of words	
golem:publicationStatus	In-Progress or Completed	
golem:publisher	Source platform	archiveofourown.org
golem:rating	Content-rating, level of sexuality/violence	Teen and Up Audiences
golem:romanticCategory	Classification for romantic relationships within the story	F/M, Gen (no rel.)
golem:socialRelationships	Social, e.g. romantic or sexual relationships between characters	Arthur/Molly Weasley
golem:series	Series the work is a part of, if any	
golem:summary	Text of the summary	
golem:title	Title	

3. Workflow

We transferred the Elasticsearch data into triple store in a series of steps. Firstly, the database was queried for stories by languages other than English. The smaller language sets were converted to triples in one step. Larger languages, such as Russian and Chinese, were processed using a batch size of 50,000 stories. The English data was queried from the database by fandom. The larger fandoms were processed in batches, while the smaller fandoms, i.e. the fandoms with few or very few stories, where queried sequentially from the database, before they were converted into triples according to the schema above.

This process has two time-consuming bottle necks: the download and the import into the Virtuoso instance. This is illustrated on the example of English fanfiction stories for the *Attack on Titan* anime in Table 2. The Elasticsearch data (jsonl format) for this fandom has a size of 4.9 GB, resulting in 60,503 triple store entries.

Table 2
Attack on Titan Example Workflow

Step	Time in Minutes
Elasticsearch Export	~20:00
Copy jsonl files	1:47
Convert to TTL format	3:20
Import to Virtuoso	9:57
Copy TTL files for backup	0:11

4. Querying the triple store

Three small case studies are presented here to demonstrate how to use the triple store and give more insights into the data contained in it. First, to know which languages are contained in the data and how many stories per language there are (stories can have more than one associated language), we can write a simple SPARQL query using COUNT. The same query can be made for different fandoms, yielding a ordered list of fandoms with the most stories (i.e. Harry Potter J.K. Rowling: 324,767, Marvel Cinematic Universe: 252,605, Supernatural: 244,182).

```
PREFIX golem: <https://golemlab.eu/graph/>
SELECT ?o (COUNT(?o) as ?oCount) WHERE
{
  ?s golem:language ?o  .
}

GROUP BY ?o
ORDER BY DESC(?oCount)
```

The results yields a list of 110 languages in total, with the top 10 by story count presented in Table 3.

Table 3
Results for the first case study, stories per languages

Language	Count
English	7,129,450
中文-普通话	448,268
Русский	148,981
Español	96,477
Français	41,006
Italiano	27,762
Português brasileiro	22,115
Bahasa Indonesia	21,605
Deutsch	17,757
Polski	15,551

Next, to find out the distribution of stories per rating (e.g. Mature or Explicit) for the fandom “Artemis Fowl - Eoin Colfer”, we can use the following query:

```
PREFIX golem: <https://golemlab.eu/graph/>
SELECT ?o (COUNT(?o) as ?oCount) WHERE
{
  ?s golem:rating ?o .
  ?s golem:fandom "Artemis Fowl - Eoin Colfer" .
}
GROUP BY ?o
ORDER BY DESC(?oCount)
```

It yields a distribution of ratings of stories within the fandom, which is normalized and illustrated in Figure 2. We can see that this particular fandom produces stories that are largely targeted at general audiences or teen and older audiences, with only flew explicit or mature stories. In contrast, the same query for the fandom BTS (a popular Korean boy band) produces a different distribution, with a larger proportion of explicit and mature stories (see Figure 3).

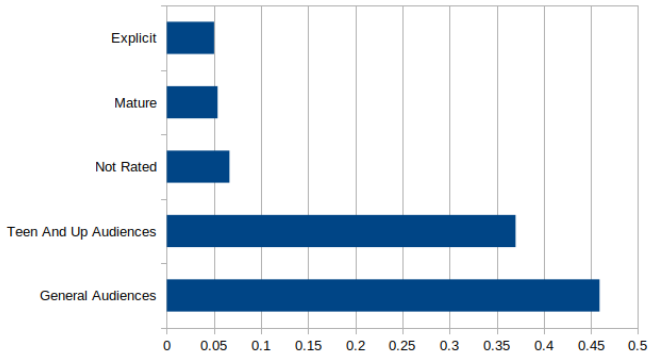


Figure 2: Distribution of Content-Ratings in Fandom “Artemis Fowl”

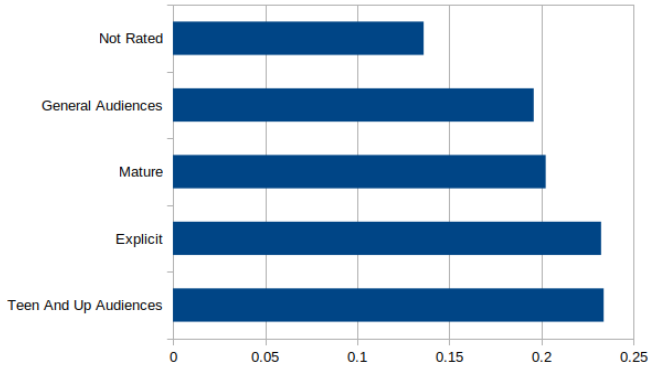


Figure 3: Distribution of Content-Ratings in Fandom “BTS”

Content-related fields are interesting for processing the fanfiction data in downstream tasks, and to derive additional semantic information on the stories. Therefore, the last example shows

how to query for a list of summaries of fanfiction stories in a specific fandom and language that are tagged with a specific keyword.

```
prefix golem: <https://golemlab.eu/graph/>
SELECT ?o WHERE
{
  ?s golem:keyword "Angst" .
  ?s golem:fandom "Attack on Titan" .
  ?s golem:language "English" .
  ?s golem:summary ?o .
}
```

The result of this query are available at https://github.com/GOLEM-lab/triple_store/.

5. Discussion

In this short paper, we present the GOLEM triple store, our effort towards a comprehensive semantic representation of fannish narratives. It provides users with manifold possibilities to study fanfiction from different viewpoints, e.g. by inspecting keywords and tags provided by the users or the distribution of romantic pairings across different fandoms. To date, the triple store contains more than 8 million stories. An overview on the statistics of the GOLEM triple store is given in Table 4.

Table 4
Triple Store Statistics

	Count
Triples	378,193,162
Stories	8,007,442
Authors	1,099,110
Fandoms	140,715*
(Stories by) Orphaned Accounts	258,736
(Stories by) Anonymous Accounts	59,007
Avg. Stories/Author	7.3

6. Future Work

The presented triple store is the first step towards a broader knowledge base for fanfiction narratives. In the short term, the triple store will be extended with additional reader response data, such as number of time users have bookmarked a story. It will further be extended from a story-centric view to a more complete data modelling based on the existing AO3 metadata, e.g. by modelling content collections according to various criteria. In the medium term, the GOLEM triple store will be extended towards a full-fledged knowledge graph of characters and events in the fanfiction domain. This includes the results of character analysis, modeling essential properties of fictional characters, i.e. physiological and psychological traits, as well

as narrative function of a character. Additionally, the full knowledge graph will also contain additional data on reader response (e.g. emotions felt, etc.) The project is currently developing a comprehensible ontology [13] for the modelling of (fan narratives), which will be aligned to relevant other ontologies, as closely as possible in order to maximize the interoperability with other relevant projects, like Wikidata and MiMoText [7].

We are additionally planning to report recent statistics of the quality of the knowledge graph (such as consistency) regularly on the project website.

Currently, the triple store only contains stories from AO3. However, we are working on including data from other sources, such as Wattpad[7] and fanfiction.net[8].

7. Acknowledgements

This work is part of the *Golem Lab: Graphs and Ontologies for Literary Evolution Models* project, a 5-year (2023-2027) research project funded by the European Commission (ERC StG).

[7] https://www.wattpad.com/
[8] fanfiction.net

References

[1] C. Fiesler, S. Morrison, A. S. Bruckman, 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, 2016, pp. 2574–2585.

[2] OECD, Derived data element, 2005. URL: https://stats.oecd.org/glossary/detail.asp?ID=5130.

[3] P. Boot, Mesotext: Digitised Emblems, Modelled Annotations and Humanities Scholarship, Amsterdam University Press, 2009.

[4] J. Jett, B. Capitanu, D. Kudeki, T. Cole, Y. Hu, P. Organisciak, T. Underwood, E. Dickson Koehl, R. Dubnicek, J. S. Downie, The HathiTrust Research Center Extracted Features Dataset (2.0), 2020. URL: https://wiki.htrc.illinois.edu/pages/viewpage.action?pageId=79069329. doi:10.13012/R2TE-C227.

[5] A. Piper, The CONLIT Dataset of Contemporary Literature, Journal of Open Humanities Data 8 (2022) 24. URL: http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.88/. doi:10.5334/johd.88, number: 0 Publisher: Ubiquity Press.

[6] S. Luoto, A. van Cranenburgh, Psycholinguistic dataset on language use in 1145 novels published in English and Dutch, Data in Brief 34 (2021) 106655. URL: https://www.sciencedirect.com/science/article/pii/S2352340920315353. doi:10.1016/j.dib.2020.106655.

[7] C. Schöch, M. Hinzmann, J. Röttgermann, K. Dietz, A. Klee, Smart Modelling for Literary History, International Journal of Humanities and Arts Computing 16 (2022) 78–93. URL: https://www.euppublishing.com/doi/10.3366/ijhac.2022.0278. doi:10.3366/ijhac.2022.0278, publisher: Edinburgh University Press.

[8] F. Fischer, I. Börner, M. Göbel, A. Hechtl, C. Kittel, C. Milling, P. Trilcke, Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama, 2019. URL: https://zenodo.org/record/4284002. doi:10.5281/zenodo.4284002, publisher: Zenodo.

[9] L. Price, Fandom, folksonomies and creativity: the case of the archive of our own (2019).

[10] M. Doerr, The CIDOC CRM, an Ontological Approach to Schema Heterogeneity, in: Y. Kalfoglou, M. Schorlemmer, A. Sheth, S. Staab, M. Uschold (Eds.), Semantic Interoperability and Integration, volume 4391 of *Dagstuhl Seminar Proceedings (DagSemProc)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2005, pp. 1–5. URL: https://drops-dev.dagstuhl.de/entities/document/10.4230/DagSemProc.04391.22. doi:10.4230/DagSemProc.04391.22.

[11] R. V. Guha, D. Brickley, S. Macbeth, Schema.org: evolution of structured data on the web, Communications of the ACM 59 (2016) 44–51.

[12] P. Riva, M. Žumer, FRBRoo, the IFLA library reference model, and now LRMoo: a circle of development, in: IFLA WLIC 2018 – Kuala Lumpur, Malaysia – Transform Libraries, Transform Societies, 2017.

[13] X. Yang, F. Pianzola, F. Pannach, The Golem Ontology: Theoretical and data-driven modelling of narrative and fiction, In Preparation (2024).