# The Changing Face of the Book Review

**Peter Boot**
Huygens Institute for the
History of the Netherlands
The Netherlands
The Hague
peter.boot@huygens.knaw.nl

**Marieke van Erp**
The Network Institute
Dept. of Computer Science
VU University Amsterdam
The Netherlands
marieke@cs.vu.nl

**Lora Aroyo**
The Network Institute
Dept. of Computer Science
VU University Amsterdam
The Netherlands
lora.aroyo@vu.nl

**Guus Schreiber**
The Network Institute
Dept. of Computer Science
VU University Amsterdam
The Netherlands
guus.schreiber@vu.nl


*WebSci 2012*, June 22–24, 2012, Evanston, Illinois, USA.
ACM 978-1-4503-1228-8

## Abstract
The Internet has drastically changed the process of literary repertoire formation - changing a closed system consisting of a few literary critics and academics to a realm of democratic rating, where each reader can voice online his or her opinions. Gathering opinions on literature from the entire spectrum, from semi-professional critics to 'ordinary' readers, and analysing and visualising these will make it possible to gain insights into the internal dynamics of online book discussion and its influences on the formation of the literary canon. With BookPulse we aim at investigating the public's influence on the canonisation process by aggregating and analysing information from different websites in order to present a more complete picture. This will help us create a personal literary Web barometer. Using as an example online discussions of the work of Dutch novelist Arnon Grunberg, we present a first analysis of the text used on a number of the review sites and analyse Twitter retweet patterns. We discuss challenges and next steps

## Keywords
Literary canon, online book reviewing, book recommendations, online book discussions

## ACM Classification Keywords
H.5.m Information interfaces and presentation (e.g., HCI) - Miscellaneous

## Introduction

With the boom of the Social Web numerous websites for book reviews also appeared. The Web has become an important publication medium for book reviews [5], though many scholars remain sceptical about the quality and impact of online reviews. Still these websites are interesting from many different perspectives in the context of book popularity building [1]. On the one hand, they can help literary scholars study people's views on literature, and provide insights into literary canonization and the formation of cultural repertoire. On the other hand, readers use these websites to stay informed about new and old books and to find out what to read next. However, it becomes more and more difficult to (1) keep track of every new website, (2) to get an overview of all book reviews, and (3) to determine the reliability and authoritativeness of the review. For Web and information scientists this creates an interesting research field: large amounts of dynamic loosely structured information across different applications that must be hooked up with existing literary information sources and personalised for access by end users. For Web archivists, the websites represent dynamic cultural heritage that is very much in danger of being lost. Finally, for the public (local) library community review sites provide both a valuable source of information about readers and a potential way to reach out to them.

BookPulse focuses on aggregating and analysing information from different websites (e.g., high-end review sites, book-based social network sites, book-oriented discussion boards) in order to present a more complete picture of the online book reviewing process. Statistical and semantic analysis of this user-generated content can uncover interesting serendipitous links between books, e.g., that the popularity of prize-winning books in a specific genre tends to have a shorter lifespan than books from a different genre. By analysing these links, we lay a foundation for personal recommendations. For public libraries, these recommendations are important because they could help patrons select books from the library stock. BookPulse will gather data from various sources in a single web archive based on open standards. As such, BookPulse will provide a testbed to explore new ways of collecting, storing and maintaining web archive content in the domain of books.

The overall goal of BookPulse is building a 'community'-influenced book reputation framework reflecting the dynamics of the current Information Age. It will develop a personalised, interactive, multimodal platform in which the success and popularity of books can be analysed in the context of how they have been discussed, reviewed, or mentioned in public Web forums. One of the computer science challenges is to link social opinions in a flexible model to identify semantically relevant relationships (e.g., 'cause' and 'effect' links along the key dimensions 'who', 'what', 'where' and 'when') between books, reviews and people in a multidimensional context.

## Research Questions

Analysing book popularity has been an active area of investigation, even before the Web. Karl Erik Rosengren measured literary fame based on book reviews focusing on two aspects (1) what books or writers make it into the hall of fame, and (2) can reviews show who's hot and who's not? [6] He showed that newspaper critics have a decisive influence on the selection of works that survive into the canon. The arrival of the Web, with weblogs, social media and book sites, provides new opportunities for ordinary readers to make themselves heard, to discover and share opinions, and to contribute to open discussions. This has deep consequences for the canonisation process of books.

However, analysing this new process is not trivial. The information is distributed and dynamic; therefore a single user can only get a limited idea about trends and reviews' networks. Aggregating relevant websites into a global, inter-

operable network brings additional knowledge. An example in the domain of movies is IMDB.com. They provide their users with STARmeter and MOVIEmeter, which take several measures of popularity, based on what and who people are looking at. To build a similar application for books, we need to crawl multiple websites, aggregate the data and identify the objects the reviews refer to: the books.

From the humanities point of view, we aim at gaining more insight in the mechanics of literary canonisation, i.e., the process through which certain literary works acquire enduring popularity and fame whereas others do not - however many copies they may sell. We focus on the influence of different aspects of New Media on this process (their appeal to a diverse audience, the possibility of aggregating data from multiple sites). On the basis of a survey, [8] found differences between omnivorous readers, who read both 'highbrow' and popular books but have less trust in traditional experts, and others. From a literary perspective, most of the research on online reviews has been into customer reviews on Amazon and has been qualitative in nature. Analysis of other review channels, especially quantitative, is almost non-existent. We consider the following research questions:

- How are processes of literary canonisation influenced by the Web, and what is the role of other media?
- How does crowd-sourcing book reputation compare to professional book reviewing in terms of impact?

From the computer science point of view, we build on existing Semantic Web technology to model social networks [2], metadata [3] and reviews in terms of events [7]. The integration and extension of these three types of data for the literature domain presents scientific challenges in terms of populating the models as well as querying and presenting the multidimensional large-scale interlinked data sets of people, books and reviews. We focus on the following research questions:

- Can we develop methods for assessing trust and reputation which are able to make connections between books, reviews, and people explicit through analysis of online book reviews and discussions?
- How can the links between books, reviews, and people be exploited to enrich user profiles and provide serendipitous recommendations?

## Exploratory investigations

To investigate the feasibility of such a system, we report on a number of initial explorations. We collected reviews about Dutch novelist and public intellectual Arnon Grunberg, one of the most prolific and highly esteemed Dutch writers of the moment. We chose Grunberg to be sure there would be a sufficient number of reviews for our analyses. We collected the reviews from a number of different sites, belonging to different site genres [1] with different characteristics in terms of required expertise, selectiveness, financial rewards, etc. These sites include bol.com (the largest online bookseller in the Netherlands), watleesjij.nu ('what are you reading now', a book-based social network site), recensieweb ('web of reviews', online review site mainly written by graduate literary students), and a number of newspapers and weeklies. The numbers for these sites are in Table 1.

| Platform | Review | Words | Reviewers | Books |
|---|---|---|---|---|
| Bol.com | 272 | 20,760 | 95 | 18 |
| Print media | 196 | 187,580 | 101 | 32 |
| Recensieweb | 14 | 11,336 | 12 | 8 |
| Watleesij.nu | 79 | 7,087 | 66 | 11 |
| Total | 561 | 226,763 | 274 | 69 |

Table 1: Reviews and reviewers data collected for the study

To contrast the reviews with a more popular means of expression, we have collected tweets related to Grunberg between January 3, 2012 and February 19, 2012 through queries that were a variation on the name 'Arnon Grunberg' and the titles

of his novels. Although Grunberg's work has been translated into other languages, we only included tweets with a Dutch language code. See Table 2 for the statistics of our Twitter collection.

| | Tweets | Words | Users |
|---|---|---|---|
| Twitter | 3,218 | 45,441 | 1,988 |

Table 2: Basic tweets statistics for the BookPulse study

## Text Analysis of Websites

With reference to the research question about changes in the processes of literary canonisation, we explore characteristics of discussions on different types of websites. Here we show one approach where we use Linguistic Inquiry and Word Count (LIWC) [4], a program that counts words in psychologically relevant semantic and grammatical categories. The dictionary used is the Dutch translation enhanced with some newer LIWC word categories and some word categories created for the purpose of this analysis. We ran LIWC on all review texts and computed average values for the four sites. The main distinction that we found was between the 'open' sites (bol.com and watleesjij.nu), and the 'closed' sites (printed media and recensieweb). Perhaps not unexpectedly, the sites differed with respect to the textual complexity of the reviews, the directness of the writing, the level of emotion displayed and the confidence in the expressed opinions: see Table 3.

The traditional review sites in all measures of textual complexity show more complexity than the sites with user-generated content. Numbers about usage of function words point in the same direction: Pennebaker argues that high usage of prepositions may indicate intellectual complexity.

The directness of the writing on the 'open' sites is seen most clearly in the high numbers of first person singular personal pronouns (I, me, my) and the high numbers of exclamation marks and of interjections (wow, cool, mmm). Conversely, the closed platforms have a high rate of 'we', showing how the writer tries to engage with his audience. On the open sites, reviewers express their feelings about the books they read much more freely than on the closed sites, giving rise to higher counts of both positive and negative emotion words. The open sites contain the words expressing certainty or confidence (sure, certain, obviously) in significantly larger numbers than the closed sites.

On a methodological note, it should be clear that the use of word count programs is not without pitfalls. The texts used for analysis should be clean and comparable. For instance, in one of the collections, the review date was included in the text. This site therefore scored significantly higher than the others in the time category. We therefore complemented category level counts with inspection of texts using a concordance program, i.e., AntConc.

## Tweets Analysis

The Twitter analysis is less straightforward than the initial analysis of the reviews from the websites. First, we determine whether a tweet is about the same 'grunberg' as the author Arnon Grunberg. Given the short length of tweets it is not clear if a tweet is about (1) Grunberg's novels, (2) his Volkskrant column, or any other Grunberg activity. For now, we focus on how the debate about this author maps out on Twitter.

To clean up, we remove all tweets of users that contain 'Grunberg'. We also do not include tweets of Arnon Grunberg himself (user @arnonyy). Furthermore, we also remove tweets of users named 'Tirza' (one of his novels). We found that queries on Grunberg's book titles were not suited for our analysis. For example, Grunberg's novels called 'Onze oom' ('Our Uncle') results in tweets of people talking about their uncles. We therefore only included tweets in which 'Grunberg' is explicitly mentioned. This resulted in 723 tweets for use in our

| Site | WC | WPS | Preps | I | we | Posemo | Negemo | certain |
|---|---|---|---|---|---|---|---|---|
| Bol.com | 76.31 | 16.99 | 8.60 | 3.02 | 0.18 | 3.60 | 3.02 | 1.31 |
| Print media | 959.13 | 21.52 | 11.68 | 0.85 | 0.42 | 1.42 | 0.85 | 0.84 |
| Recensieweb | 823.57 | 21.45 | 11.62 | 0.86 | 0.51 | 1.39 | 0.86 | 0.95 |
| Watleesij.nu | 90,873 | 17.17 | 8.49 | 4.44 | 0.16 | 3.49 | 4.44 | 1.62 |

Table 3: Average LIWC scores for the four sites. WC: word count, WPS: words per sentence, Preps: prepositions, I: first person singular, we: first person plural, Posemo: positive emotion words, Negemo: negative emotion words, certain: words expressing certainty.

analysis by 552 different users, using a total of 12,442 words or 17.21 words per tweet.

If we look at the 84 twitter users who tweeted about Grunberg at least twice, we see an interesting distribution of their types. First, we can identify at least 5 persons that are involved in debates on literature and culture professionally, e.g., fellow authors or journalists. We also identify about 7 institutions, such as the Amsterdam Library. There are also 66 private users who sometimes discuss books in their twitter feeds. The last category (6 users) we discerned in our analysis are users that only discuss book-related topics in their tweets, but their affiliation or background is unclear. Some of these users do have a considerable number of followers, and are also often re-tweeted. Although it does not seem to be real spam, the role of these users is unclear and definitely a topic for further investigation.

A first means of visualising the influence of certain users in the debate is to see who is re-tweeted often. In the centers of the clusters that we have found are either professionals in the field, or institutions. This is not so surprising, as they do have most followers. But it may indicate that influence on the twitter-verse is not so different from influence in traditional discussion forums. The full re-tweet graph can be found at http://agora.cs.vu.nl/grunberg_retweets_graphs.

## Challenges

The analyses presented are only the very first steps for our investigations. Some important challenges nevertheless emerge. An important issue is scaling up the collection of reviews. While it is comparatively easy to collect texts that mention books, it is not trivial to determine whether they classify as reviews, and if so, of what book. Getting the data cleaned up for inter-site comparison and analysis will also become more difficult when more sites and site genres are involved. For twitter discussions, obtaining clean and reliable data is an even greater challenge, but one we feel can be rewarding, as Twitter provides a timely and highly popularised peek into the current debate.

Online book reviews have up to now been mainly studied from an economic or marketing perspective. Hardly any attention has been devoted to online review sites as part of the larger literary landscape. In this contribution, we sketched a system that would provide users with personalised reading recommendations based on online book reviews published on social websites. We argued that online reviews reflect the changing processes of literary repertoire formation and that the proposed system would also allow us to study these processes. We formulated four research questions from both the humanities and computer science and reported on initial investigations into the feasibility of the proposed system. Using text analysis on the reviews from four sites, we showed

that the sites show clearly recognisable stylistic differences. Knowledge of the reviewing styles of the various sites is an important first step towards investigating deeper similarities and relations between sites. Our next aim is to create a larger archive of online reviews, which should help us investigate relations between sites and reviewers, and similarity between reviews and memberships on multiple sites.

## References


[1] Boot, P. Towards a genre analysis of online book discussion. socializing, participation and publication in the dutch booksphere. In *Selected papers of Internet Research (IR 12)* (2011).

[2] Mika, P. *Social Networks and the Semantic Web*. Springer, 2007.

[3] Miles, A., Matthews, B., Wilson, M., and Brickley, D. Skos core: simple knowledge organisation for the web. In *Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications: Vocabularies in Practice (DCMI'05)* (2005).

[4] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. *The development and psychometric properties of LIWC2007*, 2007.

[5] Pool, G. *Faint praise: the plight of book reviewing in America*. University of Missouri Press, 2007.

[6] Rosengren, K. E. Literary criticism: Future invented. *Poetics 16*, 3-4 (1987), 295–325.

[7] van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., and Schreiber, G. Design and use of the Simple Event Model (SEM). *Journal of Web Semantics 9*, 2 (July 2011), 128–136.

[8] Verboord, M. The legitimacy of book critics in the age of the internet and omnivorousness: Expert critics, internet critics and peer critics in flanders and the netherlands. *European Sociological Review 26*, 6 (2010), 623–637.