Pippa LaMacchia

Introduction to Digital Humanities

Data Biography — The Graphic Novel Corpus

February 21, 2024

The Graphic Novel Corpus stems from a German-led project by a team of researchers involved in a wide array of academic fields. Literary and cultural scholars, psychologists, and computer scientists came together to explore the under-researched field of the graphic novel. Firstly, this amalgamation of scholarship and intellectual curiosity is representative of the field of digital humanities as a whole because it takes research from a classically humanitarian sphere (the written and illustrated novel) into a digital sphere with technological experts. The collaboration between the two fields becomes representative of digital humanities itself. In this project, researchers collected 253 graphic narratives from the United States, Canada, Great Britain, and India, all written in English. This work was funded by the German Federal Ministry of Education and Research and the scholars come from the universities of Paderborn and Potsdam, both in Germany. This team of researchers approached this project because the theory and study of the illustrated graphic novel is incredibly under-studied and the group aimed to compile a source of data for public access and collaborative research.

As described by the project group, it is impossible to count the precise number of graphic novels published between the 1970s (when they began to be published in English more frequently) and 2017, the date of the research project. In an attempt to counter this issue the team drew from an array of public facing comic prizes, libraries, newspapers, and best seller lists in order to “balance popularity and prestige and to offset the biases of individual sources.” The comic data is subsequently organized into accessible charts that track a multitude of data points — information about the author, illustrator, and book itself (like country of origin, awards, pages published per-year, among many other categories). These datasets allowed the research team to participate in a multitude of conferences and contribute to greater conversations surrounding narrative storytelling and the hybridization of narrative forms.

There were several goals of this project, the first being the annotation and organization of all relevant parts of a graphic novel into a publicly accessible database. This includes several routes of categorization — as previously described, details about the author and illustrator are accessible but a particularly engaging aspect of this project are its analysis of the comic forms themselves. The team explains that “comics images are attractive targets because the structure of a comics page includes various elements (such as panels, speech balloons, captions, leading characters, and so on), the drawing of which depends on the style of the author and presents a large variability.” The project used the digitization of these works to analyze and annotate relevant visual elements of these graphic novel pages in order to compile them into visually accessible charts (one example included below). By analyzing the organization of an illustrated novel, this team enables an advancement in literary and narrative theory by solidifying and concisely visualizing a wide array of novels. In this manner, advancements can be made in the field of literary studies and image analysis.

Another aspect and goal of this project was to use eye-movement measures and eye tracking technology to further understand these multimedia narrative forms. Unfortunately, due to copyright restrictions, the scans of these graphic novels (particularly as they relate to the eye tracking portion of the study) are inaccessible to a wider public which complicates the accessibility of the project. While this major component of the project is only theoretically accessible through research articles and summaries, there are still many important questions that can be answered and addressed by what is available online.

Interactive graphs, maps and charts allow a viewer or researcher to understand the place of this Graphic Novel Corpus in the wider literary world. Differences in gender are explored, geographical range, and genre studies are available in easy-to-read and accessible charts. This allows the extension of a discourse in narrative theory and analysis of the graphic novel as a literary form. Not only is the data from their comic analysis available in this corpus, but information about this team’s work and involvement in the public sphere is clearly described. Conferences and events are clearly described, setting clear expectations about transparency and honesty with this team’s work. These articles show the actual usage, and impact of this data (instead of a theoretical application). For example, information and the slides from a masterclass at the University of Potsdam entitled “Cultural Analytics” is included as publicly available material which makes it clear for the reader exactly how this dataset was and continues to be used in an academic context.

While the compilation of this information is clearly presented in a well-designed digital space that is transparent and navigable, there are several shortcomings that are apparent. Firstly, the aforementioned impossibility of gathering *all* graphic novels published between the 1970s and 2010s mandates that there are gaps in the dataset. These researchers are transparent about this issue and attempt to circumnavigate it by utilizing a diverse array of sources, but it is impossible to fully mitigate the effects of bias if the entire range of published graphic novels isn’t accessible. Additionally, the focus on solely English language works immediately limits the study, but this limitation is ultimately understandable because if the researchers cannot understand or work with the texts they’re studying, it would create more questions about ethical study and bias. Finally, the public facing nature of this work is a way to bridge some ethical gaps in high-level academic research because so often this information is inaccessible to a general public audience but this online website/dataset is comprehensible for nearly everybody. Unfortunately, the inability to access their eye-tracking data and technology creates another chasm in the accessibility of the research but once again, this limitation is mitigated by their clarity and honesty about their analytical work.

While this dataset certainly has some ethical gaps and loopholes, it is generally accessible and incredibly comprehensible. The literary analysis that was the primary goal of this project is a success, and also shown in charts like the one included below. This dataset is now a public facing and usable resource for literary and narrative studies and the final aspect that is important to acknowledge is that this project page has not been updated since 2018 so there may have been advancements since that time within the field that are inaccessible. Otherwise, the Graphic Novel Corpus is an important early attempt at compiling data about an under-researched field for both a general audience and the academic sphere.

A close-up of a poster

Description automatically generated