Crowdsourcing annotations for comics corpora

Mihnea Tufis, Jean-Gabriel Ganascia

Crowdsourcing annotations for comics corpora

Mihnea Tufiș, Pierre and Marie Curie University, Paris 6, France
Jean-Gabriel Ganascia, Pierre and Marie Curie University, Paris 6, France

In this proposal, we address the difficulty of creating a digitized corpus by using a crowdsourced approach for annotating comic books. The resulting XML-based encodings could assist not only researchers, but publishers and collection curators equally.

The motivation for our work is three-fold. Digital Humanities scholars will be provided with an annotated corpus for conducting research relating to comics and sequential art. Curators and collectors would be provided with a structured content, which could be more easily integrated within their collections or databases. This may assist them into enlarging public or private databases of characters or comics series and enable the creation of artefacts such as comic books dictionaries, search indices and dictionaries of onomatopoeia. From a publishing perspective, the data we are collecting will allow publishers and digital comics authors to create enhanced content for a better reading experience.

Our proposal is a complementary solution to image processing approaches for identifying page/grid structures and extracting text from narration devices (i.e. different types of bubbles) [4]. We address the difficulty of automatically extracting complex page layouts, narration elements (characters, places, events, objects) or stylistic elements (frame shapes, onomatopoeia, movement lines), by engaging with the passionate "crowd" of comic books readers. Previous research has identified expertise sharing, belonging to a community and helping with a research project as strong motivating factors for crowdsourcing participants [1]. In addition, our industrial partner will incentivize participants with product vouchers for their digital comics platform.

We aggregate the answers [2] taking into account the reliability of an annotator in a given context (task difficulty, task type, annotator expertise) and the agreement between annotators [3]. We generate a quality score for each annotation, with the best of them being selected and compiled into a ready-to-use ComicsML encoding [5].

References:

[1] Dunn, S. and Hedges, M. (2012). Engaging the Crowd with Humanities Research. *Crowd-Sourcing Scoping Study*. Centre for e-Research, Dept. of Digital Humanities – King's College, London.
[2] Feng, D., Sveva, B. and Zajac, R. (2009). Acquiring High Quality Non-Expert Knowledge from On-demand Workforce. *People's Web Meets NLP-2009*, ACL (2009), 51-56.
[3] Nowak, S. and Ruger, S. (2010) How reliable are annotations via crowdsourcing? A study about inter-annotator agreement for multi-label image annotation. In *Proc. MIR-2010*, ACM, 557-566.
[4] Rigaud, C., Tsopze, N., Burie, J.-C. and Ogier, J.-M. (2011). Robust text and frame extraction from comic books. *GREC-2011*, Springer, 129-138.
[5] Walsh, J.A. (2012). Comic Book Markup Language: an Introduction and Rationale. *DHQ-6*, 1. http://www.digitalhumanities.org/dhq/vol/6/1/000117/000117.html (accessed 5 March 2016).