Open Editions Online

Jonathan Reeve, Hans Walter Gabler

15 December 2018

Scholarly editing is commonly the work of specialists. For readers and users, engaging with editions in print has in the past been essentially receptive. This need no longer be so for editions in the digital medium. Open Editions Online offers James Joyce's *Dubliners*, *A Portrait of the Artist as a Young Man* and *Ulysses* for interactive use in progressive dialog. The foundational texts are the reading texts for these works from the scholarly editions prepared by Hans Walter Gabler and his co-editors in 1993 (*Dubliners*, *Portrait*)[1] and 1984/86 (*Ulysses*)[2]. To these, our community, comprised of literature professors, Joycean enthusiasts[3], and university students, has added tens of thousands of annotations, describing the texts, and connecting them to the rich bodies of literary critical knowledge that surround them. We aim to aggregate, organize, and analyze the literary discussions related to these works, and to provide them to the public in a way that is community-owned, open access, and open source.

It is a testament to the highly allusive and intertextual nature of Joyce's works that our editions are by no means the first to attempt to reify their many allusions with hypertext. For *Ulysses* alone, there is Amanda Visconti's crowd-annotated Infinite Ulysses[4], John Hunt's Joyce Project[5], Michael Groden's "James Joyce's *Ulysses* in Hypermedia"[6], and Heyward Ehrlich's interface demonstration of "The James Joyce Text Machine,"[7] among others. We acknowledge the ingenuity of these precursors, but given their short lives as projects, we endeavor all the more carefully to choose technological systems that will ensure that our textual infrastructures are as robust and future-

[1] Joyce, Gabler, and Hettche, *Dubliners*; Joyce, Gabler, and Hettche, *A Portrait of the Artist as a Young Man*

[2] Joyce et al., *Ulysses*

[3] Besides the authors, significant contributors to this project have included Ronan Crowley (U Antwerp), Goldie Shen (Columbia U), Sophia Koh (Columbia U), Hannah House (CUNY), Fiona Liu (UC Berkeley), and the 2017 students of Sarah Cole's Columbia University course in James Joyce. More contributors are listed in the projects' GitHub repositories: https://github.com/open-editions.

[4] Visconti, "Infinite Ulysses.", archived, and no longer operating.

[5] Hunt, "The Joyce Project."

[6] See Groden, "Introduction to James Joyce's Ulysses in Hypermedia."; and Groden, "Problems of Annotation in a Digital Ulysses."

[7] Ehrlich, "James Joyce Text Machine."

proof as possible. To this end, we prioritize the creation of a well-structured, standards-compliant dataset over a user interface. This allows us to concentrate on community knowledge, rather than product creation or curation. To ensure this, we release the work under a copyleft license, thus providing the basis for reuse and remix of our work as responding users see fit[8]. Our guiding principles are decentralization, standards-compliance, and distributed versioning.

Modern digital editions need not suffer from the same limitations as their earlier counterparts, or their print versions. "Web 2.0" interactivity technologies let the user decide what he or she wants to see on a given page, allowing an edition to be simultaneously feature-rich and distraction-free[9]. As initially encountered in Open Editions Online, the texts display several of these features. Among them are *dialog attribution; text genre; language;* and *line numbers.* The categories may be toggled on or off by the reader. The *line numbers* provide elementary orientation for the digital intra-text navigation, as well as for moving between screen and book: for all three of these edited Joyce texts, the line numbers are identical throughout in the respective digital and book publications. *Text genre* and *language* help to gauge the stylistic width of the narrative. *Dialog attribution* serves to identify characters narrated as speaking and to research the intensity of their participation in the narrated events, as well as, for instance, what they are severally and together talking about.

Features we are currently developing include markup denoting cross-references, distinctive words, personal names, locations, and thoughts on the part of the narrated characters, or thematic fields on the part of the narrative texts themselves. *Cross-references* track leitmotifs and other recurring phrases, such as Bloom's "lemonyellow soap" or Stephen's "agenbite of inwit." *Distinctive words* show Joycean neologisms, according to a taxonomy that labels them as archaisms, nonstandard compounds, dialect words, and others. *Personal names* tracks the people appearing in the texts, and assigns them unique identifiers, thereby associating characters with their aliases ("Bous Stephanomenos" for Stephen; "M'Intosh" for the man in the macintosh), and linking those characters that appear in more than one work. *Locations* tracks the places mentioned in Joyce's works, differentiating between real and imagined places, and providing latitude and longitude coordinates where possible. *Thoughts* tracks the inner monologues or streams of consciousness of a character, like Stephen's musing about "Chrysostomos." *Thematic fields* comprises, for example "paralysis" in *Dubliners*, "emancipation from religion" in *Portrait*, or the "Homeric foil" in *Ulysses*.

[8]The code for the project, the contributed annotations, and other related text are released under the GNU Public License, version 3, unless already under copyright.

[9]For a good discussion of Web 2.0 technologies and *Ulysses* editions, see Marino, "*Ulysses* on Web 2.0."

Analysis

It is important to note that our markup for semantic enrichment does not merely exist to provide a richer experience for the reader, but is useful for computational analysis, as well. Since our dialog attribution markup is trivially machine-readable, for instance, it is easy to extract text files that each represent the total speech of a character in *Ulysses*. These files may then be stylistically compared to each other, using techniques of computational stylometry or forensic text analysis. Figure 1 shows the result of such an analysis, comparing the function words of the characters that speak the most, using principal component analysis[10]. The axes represent the first and second principal components, and the proximity of points in this space represents the stylistic similarity of those characters' utterances. There are clear affinities here between the co-protagonists Stephen and Bloom, who cluster very closely together; similarly among the journalists and literary men Eglinton, Crawford, Hynes, and MacHugh.

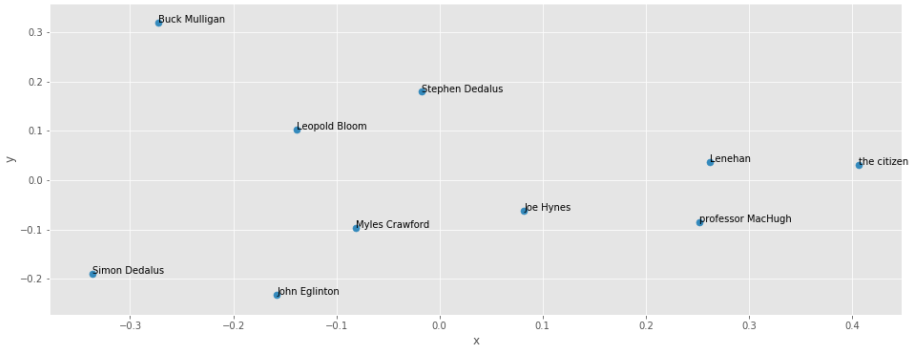

Figure 1: Similarity analysis of character style in *Ulysses*

Figure 2 shows another type of analysis: a quantification of the titles of mentioned works, arranged by episode and media type. With this visualization, it is possible to see at a glance which episode is most occupied with newspapers ("Aeolus"), books ("Ithaca"), or songs ("Sirens"). While this doesn't reveal anything surprising—critics have long discussed the intermediality of *Ulysses*—it is an example of how ad hoc schematization of the novel is possible with machine-readable markup in TEI XML.

Many more varieties of analysis are made possible by our markup, both computational and traditional. Figure 3 shows a subset of some of the places mentioned in *A Portrait of the Artist as a Young Man*, as encoded by our collaborator Goldie Shen, and visualized by Hannah House. Darker points indicate later positions in the narrative time of the novel. Again, this does not reveal anything

[10]For the use of function words as stylistic indicators, see Damerau, "The Use of Function Word Frequencies as Indicators of Style."; for a seminal study using this technique, see Burrows and Hassall, "Anna Boleyn and the Authenticity of Fielding's Feminine Narratives."

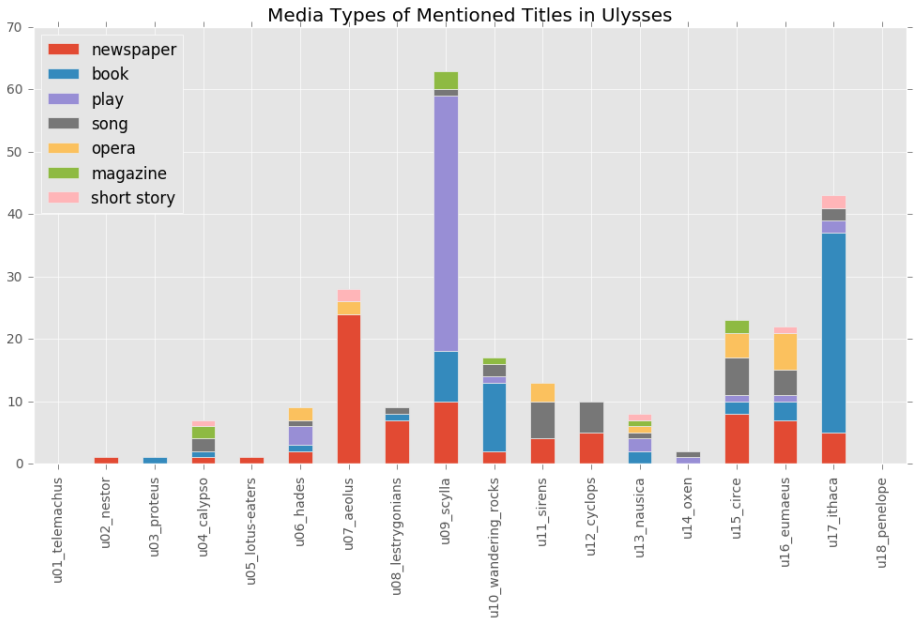

Figure 2: Media Types of Mentioned Titles in *Ulysses*

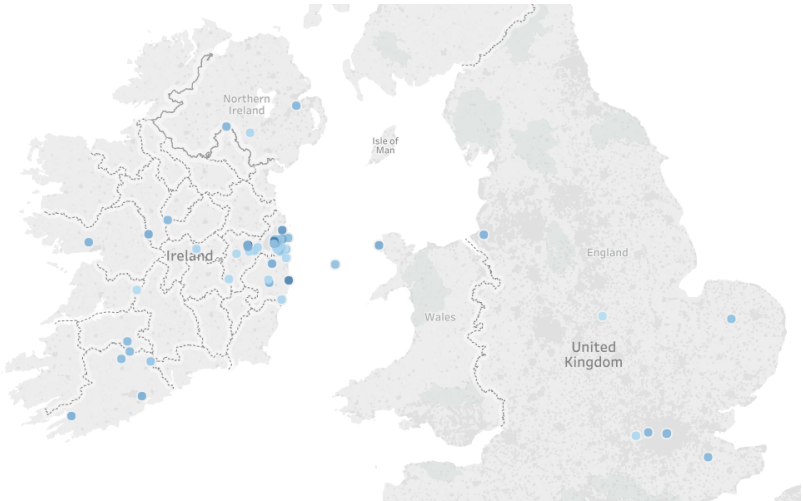

Figure 3: Map of Places Mentioned in *A Portrait of the Artist as a Young Man*

unknown to critics, and in fact it is clear from the political divisions here that this is an anachronistic projection, but this nonetheless allows one to see at a glance the geographic movement of the novel's *Bildung*: toward Dublin.

The sources of knowledge we aggregate are many and varied: entries from *Thom's Dublin Directory, 1904* recently made available by the web resource Joyce Tools[11]; period maps of Dublin; and anything else we find in the public domain. We are joining forces with the the Joyce Word Dictionary[12], to automatically populate the lexicon with words we mark as distinct, and to integrate the words' descriptions from that lexicon into our text. Soon, we will use text reuse detection technology, originally designed to detect plagiarism, to find quotations from these Joyce works beyond even those critical articles that cite them. This will allow us to say, for instance, what the most-quoted Joycean sentence is, what is the least quoted, and how these patterns change over time[13]. More pertinently, even, text reuse technology will prove usable to analyze *Ulysses* in terms of Joyce's assimilation of extrinsic reading matter as well as of his recurrent reuse of language and text from the book itself.

Technologies

Open Editions leverages two well-established technologies: we mark up text using TEI XML (Text Encoding Initiative eXtensible Markup Language[14]), and track changes with the distributed version control system Git[15]. TEI XML is the encoding format used by approximately three thousand texts on the Oxford Text Archive[16]. Whereas traditional text encoding only describes how a textual unit *looks*—as in the HTML `<em>Introibo ad altare Dei</em>`, which uses the `<em>` tag to indicate emphasis—we describe rather what it *is* and *does*, as `<foreign xml:lang="lat">Introibo ad altare Dei</foreign>`, which shows that it's a foreign expression in the Latin language. Future work will show that this is Latin as deployed in Church service, and will provide a external link to a text of the mass.

The distributed version control system we use, Git, has long been in widespread use in software development, but is rarely used for textual editing. When applied to textual editing workflows, however, Git, along with its related project management service, GitHub, becomes a powerful platform that enables collab-

[11] Gunn, "Joyce Tools."

[12] Chenier and O'Sullivan, "Regarding the Joyce Word Dictionary."

[13] For a similar project, see *Middlemarch Critical Histories*, a work currently in progress at the Literary Modeling and Visualization Lab at Columbia University: Reeve, Terlunen, and Eckert, "Frequently Cited Passages Across Time."

[14] See burnard1988report;. An earlier version of this encoding system is described in Burnard and Sperberg-McQueen, *Guidelines for Electronic Text Encoding and Interchange*.

[15] For a general introduction to Git, see Loeliger and McCullough, *Version Control with Git*; for a more humanities-oriented tutorial, see Van Strien, "An Introduction to Version Control Using Github Desktop."

[16] Hosted at http://ota.ox.ac.uk; see also Proud, *The Oxford Text Archive*.

oration, promotes the creation of editorial metadata (think: a scholarly "track changes"), and keeps every version of the text. Among other benefits, this provides a radical decentering of the power structure of a traditional editorial committee, aligning it more with open-source software projects. Any interested party may contribute markup or code to the project. While we, the project coordinators, seek to maintain a core or parent Open Edition version and anticipate selecting incoming contributions for inclusion in *our* version, parallel versions may be created and published in seconds, for private use, or published under open access conditions, each reflecting their creators' editorial and analytic interpretive choices. At the time of writing, our editions have seen a total of around twenty contributors, and this number continues to grow. Additionally, since Git requires detailed descriptions of each change or set of changes, this provides us with a mechanism by which to document every change we make, to revert changes as necessary, and to collect sets of changes in "branches." This builds on the already successful model of Wiki-based sites like Wikipedia, but adds greater editorial accountability and a well-defined process of peer review.

Call for Contributions; Future Work

Like Wikipedia, or other crowdsourced knowledge bases, Open Editions is only as good as its volunteer contributions. To that end, we are actively seeking contributions from Joyceans. No prior technological experience is required, and instructions are simple and easy to follow. Contribution simply entails creating an account, and surrounding a passage of text with a tag that indicates its function, as with the `<foreign>` example above. For more, please see the contribution instructions on our website[17].

In the coming years, our project will likely expand beyond the works of James Joyce. We are currently building a system that will take as input any TEI XML document, and output an interactive reading edition for it. We hope that this will transform the way we think about and interact with the works of James Joyce, as well as other literary texts.

References

Burnard, Lou, and CM Sperberg-McQueen. *Guidelines for Electronic Text Encoding and Interchange*. Association for Computers and the Humanities, Association for Computational Linguistics, Association for Literary and Linguistic Computing, 1994.

Burrows, John Frederick, and Anthony J Hassall. "Anna Boleyn and the Authenticity of Fielding's Feminine Narratives." *Eighteenth-Century Studies* 21, no. 4 (1988): 427–53.

[17] http://open-editions.org/contributing

Chenier, Natasha R, and James O'Sullivan. "Regarding the Joyce Word Dictionary." *James Joyce Quarterly* 53, no. 1 (2015): 17–21.

Damerau, Fred J. "The Use of Function Word Frequencies as Indicators of Style." *Computers and the Humanities* 9, no. 6 (1975): 271.

Ehrlich, Heyward. "James Joyce Text Machine." Digital edition, 2001. http://www.joyceproject.com/.

Groden, Michael. "Introduction to James Joyce's Ulysses in Hypermedia." *Journal of Modern Literature* 24, no. 3 (2001): 359–62.

———. "Problems of Annotation in a Digital Ulysses." *JoyceMedia: James Joyce, Hypermedia and Textual Genetics*, 2004, 116–32.

Gunn, Ian. "Joyce Tools," 2018. http://www.riverrun.org.uk/joycetools.html.

Hunt, John. "The Joyce Project." Digital edition, 2015. http://www.joyceproject.com/.

Joyce, J., H.W. Gabler, and W. Hettche. *A Portrait of the Artist as a Young Man*. Case Studies in Contemporary Criticism. Garland Pub., 1993.

———. *Dubliners*. Garland Pub., 1993.

Joyce, J., H.W. Gabler, W. Steppe, and C. Melchior. *Ulysses*. Literature (Vintage Books). Vintage Books, 1986.

Loeliger, J., and M. McCullough. *Version Control with Git: Powerful Tools and Techniques for Collaborative Software Development*. O'Reilly Media, 2012.

Marino, Mark C. "*Ulysses* on Web 2.0: Towards a Hypermedia Parallax Engine." *James Joyce Quarterly* 44, no. 3 (2007): 475–99.

Proud, Judith K. *The Oxford Text Archive*. Vol. 5985. British Library, Research; Development Department, 1989.

Reeve, Jonathan Pearce, Milan Terlunen, and Sierra Eckert. "Frequently Cited Passages Across Time: New Methods for Studying the Critical Reception of Texts." In *Digital Humanities Conference Abstracts*, 2017.

Van Strien, Daniel. "An Introduction to Version Control Using Github Desktop." *The Programming Historian*, 2016. https://programminghistorian.org/en/lessons/getting-started-with-github-desktop.

Visconti, Amanda. "Infinite Ulysses." Digital edition, 2013. http://infiniteulysses.com/.