Building an Edition from Journal Articles

Wandel, Wert, und Wirkung von Editionen, 2023-09-20

Charles H. Pence

@pence@scholar.social

UCLouvain

Institut supérieur de philosophie (ISP)

ACADÉMIE ROYALE DE BELGIQUE 1772

The Sciveyor Project:
A Post-Mortem

Outline

1. What matters, and to whom?
   1.1 Completeness
   1.2 Versioning
   1.3 Pragmatics
2. Looking outward

The take-home: An “edition” of journal articles is a slippery, complex, contextual, local thing—one that might not exist.

My larger goal for today: Try to draw some morals from the Sciveyor story about:

1. **Digital editions in general**
2. What some of Sciveyor's failures can tell us about **potentially overlooked aspects of edition curation** (or, at least, aspects that *I* had overlooked!)

What Matters, and to Whom?

Completeness

Completeness

There is an extremely natural tendency when running a digital analysis to want to run it **against a complete corpus**

Completeness

Of course, there's no such thing as a **complete edition.**
Every boundary-setting judgment is a subjective one,
even if we know how to defend them well.

But things get worse for journal articles.

Completeness

- Copyright and moving walls
- Journal scale
- Generalist vs. specialist journals
- Historical OCR quality

Versioning

Versioning

An edition, properly speaking, lets us **version** our interactions with the text. This is **absolutely vital** for reproducibility, whether of digital or of analog research!

Versioning

But for always-online digital editions, especially those where we might want to make frequent updates (new content, revised translations, new processing steps, etc.), how can we guarantee stable texts?

Must “online once” = “online forever?”

Versioning

Again, this problem is magnified for journal articles:

- Need to add most recent articles
- Improvements to OCR
- Renegotiations of copyright agreements

Pragmatics

Pragmatics

How ought we take user interests into account? How **divergent** do those user interests wind up being?

Pragmatics: An Illustration

I received an e-mail a few years ago, inquiring about a few apparently broken analyses. A biologist was searching for the gene SH-SY5Y in articles in Sciveyor, and nothing worked.

Pragmatics: An Illustration

But: I’d never considered analyzing tokens like that – separated by a hyphen, composed of letters and numbers. Does my tokenizer even produce them as analyzable tokens? Can I search for them? I had no idea! It was a **pragmatic user need** that I had never thought about.

Pragmatics and Completeness

Another aspect relates to completeness: every user **has their own Reviewer 2** that they need to satisfy. Will the corpus always give them the materials that they need to do that? How can we know in advance? And what can we do if the answer is no?

Looking Outward

Summing up

1. What does it mean for a corpus to be complete?
2. The importance of versioning our interactions with text
3. How should we incorporate end-user needs?

An edition of journal articles?

So, can we actually produce something that feels like an "edition" of journal articles?

An edition of journal articles?

So, can we actually produce something that feels like an “edition” of journal articles?

In the end… **I’m not so sure.**

If I started today…

Because of pragmatic concerns, these kinds of corpora seem to be perennially **contextual, local, and smaller scale** than I had first envisioned.

How can we respond to that need in a way that is, or is built out of, a large, publicly accessible resource? And how could we get that resource past Reviewer 2?

If I started today…

Perhaps the right play is to think about a kind of **meta-edition** framework here: a system that lets users **generate** the bodies of text that they need.

What kinds of **guarantees** would that system need to offer to users, to reviewers, and to the scholarly public?

If I started today…

At the very least, **Sciveyor failed to:**

- Expose information about the kinds of subjective decisions that we made
- Permit permanent reference to texts used for an analysis
- Help users motivate the virtues of their corpus

Questions?

charles@charlespence.net
https://pencelab.be
@pence@scholar.social

PENCE LAB

fnrs
LA LIBERTÉ DE CHERCHER