

# Using the Hypothes.is web annotation tool for neologism collection

Erin McKean  
Wordnik.com  
8 May 2019

# What's Hypothes.is?

- A US-based 501(c)3 on a mission to build a “conversation layer” for the web
- Funded through Mellon, Knight, Sloan + other major foundations
- Annotation standard is based on the W3C Web Annotation Working Group
- More than 5M annotations as of March 2019

# What's Wordnik?

- A US-based 501(c)3 nonprofit
- Mission to 'collect and share all the words of English'
- Founded 2008, became nonprofit 2014
- Wordnik.com website + developer API
- Information about >8M words

# Hypothes.is Technical Overview

- Annotation service that stores, searches, and displays annotations, manages users and groups, and manages browser clients
- Annotation viewer that runs in the browser
- Bookmarklet for marking annotations in the browser
- API service

# Setup

- Create a Hypothes.is account (free)
- Create a Hypothes.is group
  - Currently only private groups are self-service
  - Restricted and open groups are in Publisher beta
- Invite users to group
  - Users must have (free) Hypothes.is account

# Wordnik Program

- Set up Hypothes.is group in 2017 as a possible adjunct to a class in lexicography, but did not see much use
- Began recruiting active users of the Wordnik site as readers in Q3 2018

# Citation Annotation Process: sign up

We're currently beta-testing a way for readers to add citations from the web to Wordnik. To join, please first sign up for a [hypothes.is](https://www.hypothes.is) account and then fill out this form.

\* Required

**Email address \***

Your email

---

**What is your Wordnik username?**

# Citation Annotation Process: join group

## Group invitation

You've been invited to join the annotation group:

### **Wordnik-Friends**

Friends of Wordnik what like to find words

**Log in to join Wordnik-Friends**



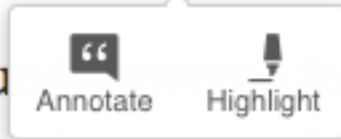
# Citation Annotation Process: install Hypothes.is bookmarklet in browser



# Citation Annotation Process: markup

While people living in the region knew of the termite mounds, few outsiders did. The expanse of the termites' construction were hidden by scrubby forest known as caatinga.

“That's why they were u r so long,” Dr. Funch said. “You cannot see them in the native vegetation. And not many scientists pass this way.”



# Citation Annotation Process: markup

The screenshot shows the Wordnik-Friends annotation interface. At the top, the group name "Wordnik-Friends" is displayed. Below it, the title "esperluette" is shown with a duration of "33 secs". The user "Wordnik-Friends" is listed as the author. The main text of the annotation is a quote: "The expanse of the termites' construction were hidden by scrubby forest known as caatinga." Below the text is a rich text editor with a toolbar containing icons for bold, italic, quote, link, image, sum, list, and list, along with a "Preview" button. At the bottom, there is a tag input field containing the tags "hw-caatinga" and "wordnik", followed by "Add tags...". Below the tag field are two buttons: "Post to Wordnik-Friends" and "Cancel".

# Hypothes.is API: selected response fields

**"updated"**: "2018-11-27T22:46:28.917692+00:00",

**"exact"**: "I've coined a new term for the gallows humor that my generation indulges in because we have an overheating planet, a dim political future, a crushing economy, and a real avocado toast problem:Millennihilism",

**"tags"**: [ "wordnik", "hw-millennihilism" ],

**"created"**: "2018-11-27T22:46:28.917692+00:00",

**"uri"**: "http://copperbadge.tumblr.com/post/161095955784/ive-coined-a-new-term-for-the-gallows-humor-that",

**"user"**: "acct:Purplatypus@hypothes.is",

# Known Problems in Citation Selection

- Over-citing for rare words
- Missing/incomplete attributions
- Storage
- Recruitment/motivation of readers
- Setup/management costs

# Known Problems in Citation Selection

- *Over-citing for rare words*
  - Not a problem for Wordnik's model
- *Missing/incomplete attributions*
  - Source urls automatically produced by Hypothes.is
  - May still be an issue for poorly-coded web pages (missing metadata)

# Known Problems in Citation Selection

- *Storage*
  - Kept on Hypothes.is servers: can run own server/backups
- *Recruitment/motivation of readers*
  - Still an issue ...
- *Setup/management costs*
  - Editorial time spent in creating instructions, answering questions
  - Engineering time in writing connectors to Hypothes.is API, setting up db
  - Review process setup (editorial and technical)

# Process Issues

- Bookmarklet can balk at some web pages, esp if iframes are blocked
- Citation readers must conform to tagging instructions
- Problems annotating PDFs with no text layers



# Possible Risks

- Weaponization by bad actors
  - Spammers, ax-grinders
- Hypothes.is shutdown or software deprecation
  - Server is open source; can run own server instance
- Overgamification
  - Leaderboard obsession

# What's Next?

- Citations gathered through this process will begin appearing on the site in Q3 2019
- Plan to mark user-gathered citations with an icon or badge that links to a “what’s this?” page that includes signup for new volunteers
- User leaderboard to show recent citations, citation totals
- Simplification of tag schema

# What's Next? Pt. 2

- Recruitment of readers interested in specific subject areas
- New features in Hypothes.is API allow for retrieval of all annotations by url, permitting selection of random (but human-marked) annotations

# Conclusions

- Hypothes.is is a relatively simple, cost-effective solution for marking and retrieving citations from web pages .... if:
  - Over-citation of rare and nonce words is not an issue
  - Text-type balance is not important (as citations are only from web sources)
  - Sufficient resources are available to manage readers and create API connectors

# Questions?

- Links:
  - [web.hypothes.is](http://web.hypothes.is)
  - [web.hypothes.is/developers](http://web.hypothes.is/developers)
  - [wordnik.com](http://wordnik.com)
  - [developer.wordnik.com](http://developer.wordnik.com)