

Text

Text



What is "Text"?

In scientific computing in general, and linguistics in particular, we're going to be working with text a lot, specifically "raw text." Code we write is going to be written in raw text, the best way to do textual analysis is with raw text, and even some of the writing we'll do will be in raw text.

But what is "raw text"? We write and read a lot every day, whether it's text messages, social media posts, emails, or papers for course, but these texts are usually not written in or viewed in "raw" form. There is a *lot* more information underlying these formats than is immediately made visible to us.



Discussion

If we write just a simple single sentence in a Google Doc, what other kinds of information is present in the document beyond the letters and punctuation of the text?



Markup and Metadata


Most of the information beyond the raw text can be classified as "markup" and "metadata" that provide instruction to the application (like the web browser, app, or word processor) to *display* text in a particular way. You've probably already encountered this if you've ever copy-pasted text into Microsoft Word and been asked to choose whether to "Keep Source Formatting", "Match Destination Formatting" or "Keep Text Only".

Most of the information beyond the raw text can be classified as "markup" and "metadata" that provide instruction to the application (like the web browser, app, or word processor) to *display* text in a particular way.



The *amount* of markup and metadata for any piece of text can be huge. Take, for example, this simple tweet of mine.

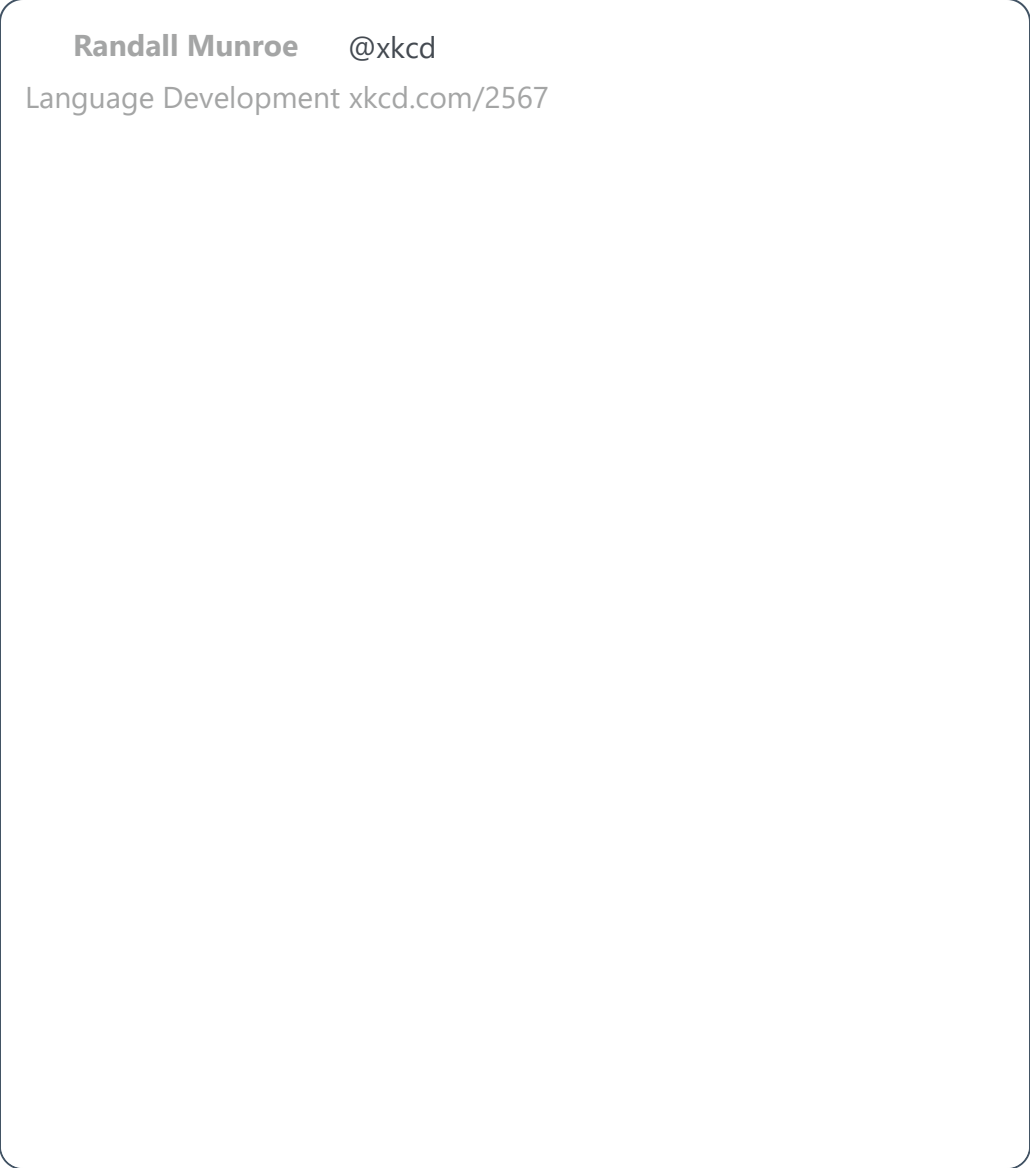
The Fruehwald
@JoFrhwld · [Follow](#)







i'm teaching Child as a Linguistic Historian on Monday, so yes, good

Randall Munroe @xkcd

Language Development xkcd.com/2567



3:09 PM · Jan 12, 2022 

 **40**  **Reply**  **Copy link**

[Read more on X](#)

I downloaded all of the data that is packaged up in this tweet, which winds up looking like this:

```

1  [
2    {
3      "user_id":"14730367",
4      "status_id":"1481357767550066688",
5      "created_at":"2022-01-12 20:09:48",
6      "screen_name":"JoFrhwld",
7      "text":"i'm teaching Child as a Linguistic Historian on Monday, so yes, good https://t.co/gFhE9WzMNZ",
8      "source":"Twitter for iPhone",
9      "display_text_width":68,
10     "is_quote":true,
11     "is_retweet":false,
12     "favorite_count":57,
13     "retweet_count":0,
14     "hashtags":[null],
15     "symbols":[null],
16     "urls_url":["twitter.com/xkcd/status/14..."],
17     "urls_t.co":["https://t.co/gFhE9WzMNZ"],
18     "urls_expanded_url":["https://twitter.com/xkcd/status/1481328702734716933"],
19     "media_url":[null],
20     "media_t.co":[null],
21     "media_expanded_url":[null],
22     "media_type":[null],
23     "ext_media_url":[null],
24     "ext_media_t.co":[null],
25     "ext_media_expanded_url":[null],
26     "mentions_user_id":[null],
27     "mentions_screen_name":[null],
28     "lang":"en",
29     "quoted_status_id":"1481328702734716933",
30     "quoted_text":"Language Development https://t.co/yw1x7ct3JD https://t.co/XoHK3CMAgt",
31     "quoted_created_at":"2022-01-12 18:14:19",
32     "quoted_source":"Twitter for iPhone",
33     "quoted_favorite_count":5356,
34     "quoted_retweet_count":770,
35     "quoted_user_id":"21146468",
36     "quoted_screen_name":"xkcd",
37     "quoted_name":"Randall Munroe",
38     "quoted_followers_count":186732,
39     "quoted_friends_count":1,
40     "quoted_statuses_count":351,
41     "quoted_location":"","
42     "quoted_description":"I draw the comic xkcd",
43     "quoted_verified":true,
44     "geo_coords":["NA","NA"],
45     "coords_coords":["NA","NA"],
46     "bbox_coords":["NA", "NA", "NA", "NA", "NA", "NA", "NA"],
47     "status_url":"https://twitter.com/JoFrhwld/status/1481357767550066688",
48     "name":"The Fruehwald",
49     "location":"Lexington, KY",
50     "description":"I'm a linguist. Not a real sir. And I enjoy knitting. Asst Prof @ University of Kentucky he/him",
51     "url":"https://t.co/v6483GMM7y",
52     "protected":false,
53     "followers_count":6235,
54     "friends_count":1877,
55     "listed_count":115,
56     "statuses_count":28157,
57     "favourites_count":34699,
58     "account_created_at":"2008-05-11 03:03:39",
59     "verified":false,
60     "profile_url":"https://t.co/v6483GMM7y",
61     "profile_expanded_url":"http://jofrhwld.github.io/",
62     "profile_banner_url":"https://pbs.twimg.com/profile_banners/14730367/1599769471",
63     "profile_background_url":"http://abs.twimg.com/images/themes/theme7/bg.gif",
64     "profile_image_url":"http://pbs.twimg.com/profile_images/1234882113289310210/32qkUIbT_normal.jpg"
65   }
66 ]

```

Only one line out of all of that information is the actual text of my tweet!

The specialized instruction for how to display text is called "markup." Here's what the following sentence looks like "rendered" in html, and after that how it looks in raw text.

In this sentence, *this span of text* is italicized.

```
In this sentence, <i>this span of text</i> is italicized.
```

In the "rendered" version, the text just appears in italics. But in the raw text version, we can see the behind-the-scenes instruction to italicize the text. The `<i>`, or opening "tag", instructs your browser to start displaying the following text in italics, and the `</i>`, or closing "tag" tells the browser where the italics stop.

There are lots of different kinds of markup languages out there, each of which have their own way of rendering italics text.

LaTeX

```
In this sentence, \textit{this span of text} is italicized.
```

Markdown

```
In this sentence, *this span of text* is italicized.
```



Some fair questions

Why do we need to understand the difference between raw text and rendered text?

First and foremost, the scientific programming languages we're going to be working with this semester are written in raw-text, and can only see raw-text. For example, if we point python at a website to analyze the text on the website, it's going to be looking at the text with all of the tags like `<i></i>`, not the nicely rendered text that we look at.

Also, sometimes it's easier and prettier to create diagrams, tables, and other kinds of things for our linguistics papers in markup languages directly.

Why is the font so ugly in the raw text?

```
It's a fair question to ask why we have to look at raw text in this weird font.
```

This font is called "fixed-width" because every single character in the font has the same width. Characters that would be slender in most other fonts, like "I", or a period take the same width on the screen. Variable width fonts are aesthetically more pleasing, and sometimes easier to read, but it's easier to visually line up text vertically with fixed width fonts, which is often very useful when writing a computer program. Compare how these two blocks of text look different in the two kinds of font.

p.hello()
i.hello()
m.hello()

```
p.hello()  
i.hello()  
m.hello()
```

How do I write raw text?

There are lots of options in "text editors". Many programming languages have their own "Integrated Development Environments" or IDEs with built in text editors.

You can check out the [Text Editors \(https://uk.instructure.com/courses/2051722/pages/text-editors\)](https://uk.instructure.com/courses/2051722/pages/text-editors) page for a list and links to some nice options.