Social Data Science Base Camp

Digital Data Structures

Friedolin Merhout
MSc Social Data Science
September 20, 2021

UNIVERSITY OF COPENHAGEN

Outline for today

1. Welcome back and check-in
2. Lecture
    1. Data Formatting
    2. Hypertext Markup Language (HTML)
    3. Cascading Style Sheets (CSS)
    4. JavaScript Object Notation (JSON)
3. Demo
    1. Hands on “W38Mon-Demo-DigitalDataStructures.ipynb”
4. Review Survey
5. Exercise Preview

Welcome back and check-in

Image Source

- Recap
  - (Conditional) flow control
  - Naming Practices
  - Containers / Data Structures
    - Access, order, duplication, mutability
- Outlook
  - Wednesday: Loops / (Repetition) flow control
  - Next Monday: Review and Building Bigger Programs

More Structures I

Digital Data Structures

Building on slides by Ulf Aslak

Wake Up Exercise: Sort that file

File Name	Data Type
“description.txt”	
”rows.csv”	
“table.html”	
”observations.json”	
”data.xlsx”	

Wake Up Exercise: Sort that file

File Name	Data Type
“description.txt”	Non-tabular
”rows.csv”	
“table.html”	
”observations.json”	
”data.xlsx”	

Wake Up Exercise: Sort that file

File Name	Data Type
“description.txt”	Non-tabular
”rows.csv”	Tabular
“table.html”	
”observations.json”	
”data.xlsx”	

Wake Up Exercise: Sort that file

File Name	Data Type
“description.txt”	Non-tabular
”rows.csv”	Tabular
“table.html”	Non-tabular
”observations.json”	
”data.xlsx”	

Wake Up Exercise: Sort that file

File Name	Data Type
“description.txt”	Non-tabular
”rows.csv”	Tabular
“table.html”	Non-tabular
”observations.json”	Non-tabular
”data.xlsx”	

Wake Up Exercise: Sort that file

File Name	Data Type
“description.txt”	Non-tabular
”rows.csv”	Tabular
“table.html”	Non-tabular
”observations.json”	Non-tabular
”data.xlsx”	Tabular

What does this mean?

Data Formatting

- Gist:
  - Varying nature and application of data leads to wide variation in representation and storage formats

- Three main types of formats
  - Tabular (CSV, XLSX)
  - Non-tabular (JSON, XML)
  - Associative (*Networks*)

Data Formatting

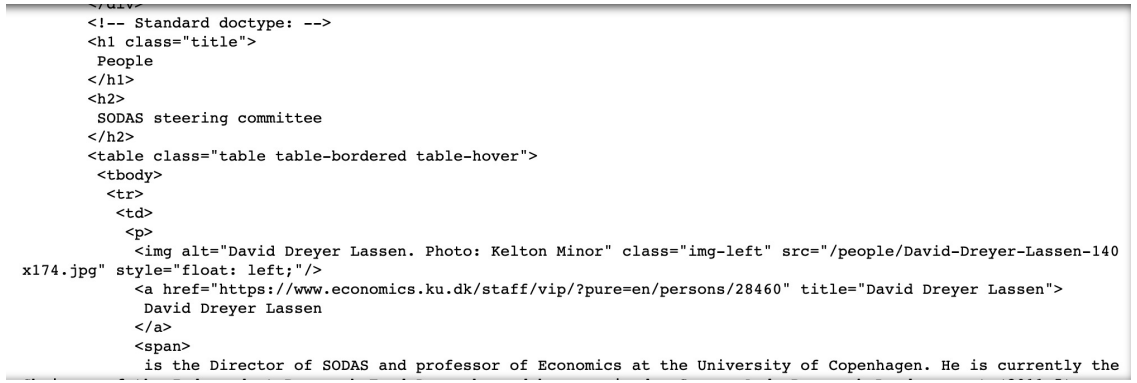
- Gist:
  - Varying nature and application of data leads to wide variation in representation and storage formats
- Three main types of formats
  - **Tabular (CSV, XLSX)**
    - Structure: rows and columns
    - Common terminology: rows/observations/cases, columns/vectors/variables

	description	role
0	David Dreyer Lassen is the Director of SODAS a...	SODAS steering committee
1	Morten Axel Pedersen is Deputy Director of SOD...	SODAS steering committee
2	Rebecca Adler-Nissen is Professor in Political...	SODAS steering committee
3	Sune Lehmann is a Professor of Complexity and ...	SODAS steering committee
4	Anders Blok is Associate Professor in Sociolog...	SODAS steering committee
5	Søren Kyllingsbæk is Professor in Cognitive Ps...	SODAS steering committee
6	Robert Böhm is a Professor of Applied Social P...	SODAS steering committee

Data Formatting

- Gist:
  - Varying nature and application of data leads to wide variation in representation and storage formats

- Three main types of formats
  - **Non-tabular (JSON, XML)**
    - Structure: varying but often tree-like (stem-branches-leaves) or hierarchical
    - Common terminology: keys/attribute, values



```
<!-- Standard doctype: -->
<h1 class="title">
 People
</h1>
<h2>
 SODAS steering committee
</h2>
<table class="table table-bordered table-hover">
 <tbody>
  <tr>
   <td>
    <p>
     <img alt="David Dreyer Lassen. Photo: Kelton Minor" class="img-left" src="/people/David-Dreyer-Lassen-140
x174.jpg" style="float: left;"/>
     <a href="https://www.economics.ku.dk/staff/vip/?pure=en/persons/28460" title="David Dreyer Lassen">
      David Dreyer Lassen
     </a>
     <span>
      is the Director of SODAS and professor of Economics at the University of Copenhagen. He is currently the
```

Data Formatting

- Gist:
  - Varying nature and application of data leads to wide variation in representation and storage formats

- Three main types of formats
  - **Associative (*Networks*)**
    - Structure: network (connected observations)
    - Common terminology: nodes/vertices, edge/tie/link

Image Source

Today's focus: non-tabular data in the wild

Digital Data Structures

HyperText Markup Language

HyperText Markup Language

- **Web pages** are built up and stored as plain text, encoding hypertext controlling appearance and functionality of pages

- To ensure proper presentation of pages, links, lists, etc., requires **common format** for marking up text: That's **HTML**.

Image Source

HTML Fundamentals

- HyperText Markup Language is a hierarchical format using **pre-defined tags** to structure text and media content

- Existing tags include indicators for headings `<h1>`, image embedders `<img>`, and hyperlinks `<a>` and many more

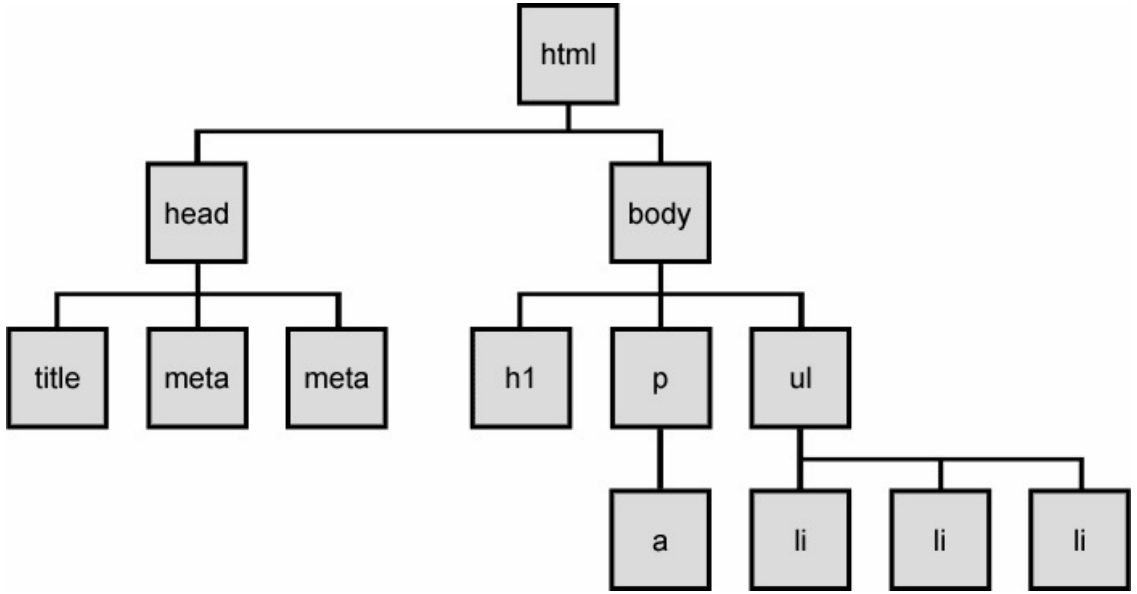


Image Source

HTML Tags I

- Some tags **wrap around** content and need to be closed, e.g.
  - `<h1>This is a heading</h1>`
- Others **contain** content and can stand by themselves, e.g.
  - `<img src="cute-cat.jpg">`

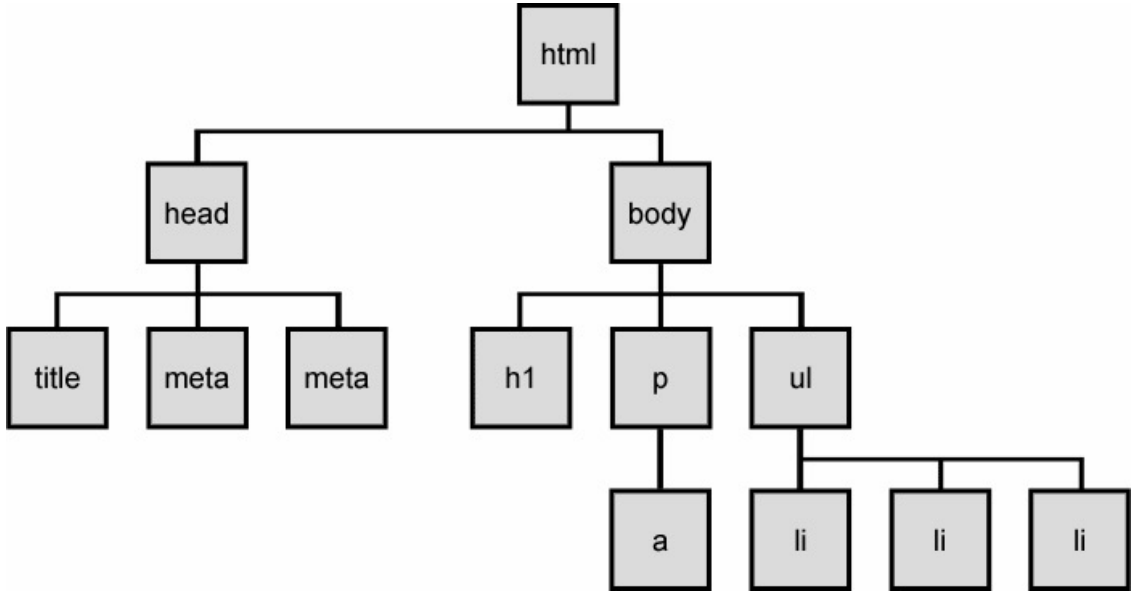


Image Source

HTML Tags II

- Tags can have **attributes** with additional information, e.g.

```
<img src="cute-cat.jpg">
<a
href="https://fmerhout.github.io">a link</a>
```

- One fundamental tag is `<div>`, defining sections containing distinct content

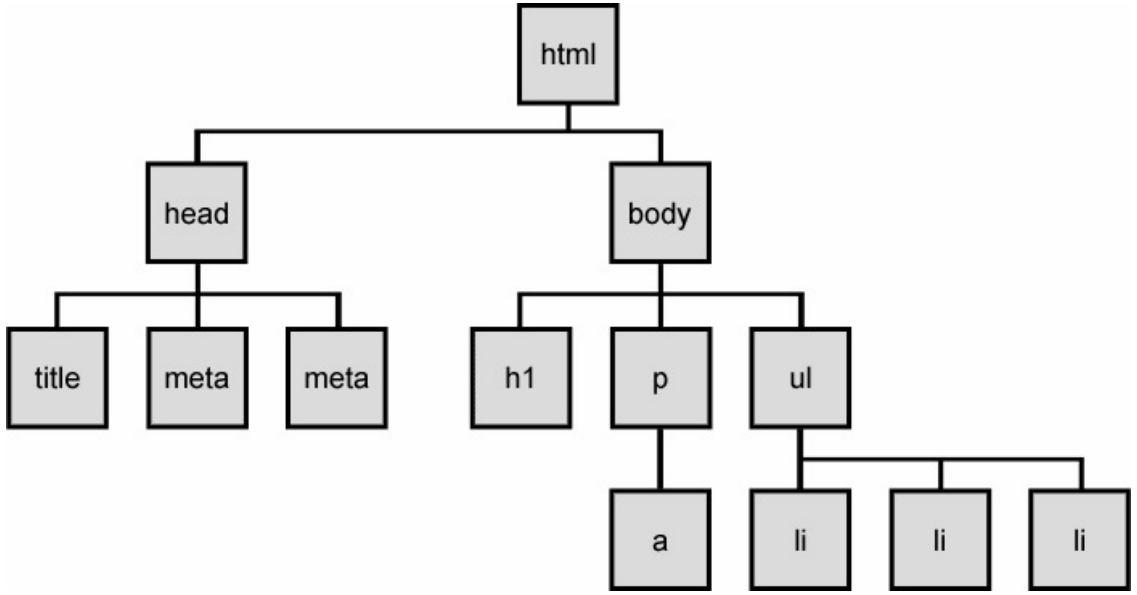


Image Source

HTML in Practice

- Browser create internal representation of plain text HTML called **Document Object Model**
- Modern websites build on frameworks allowing users to interact with it, **dynamically** loading or changing content

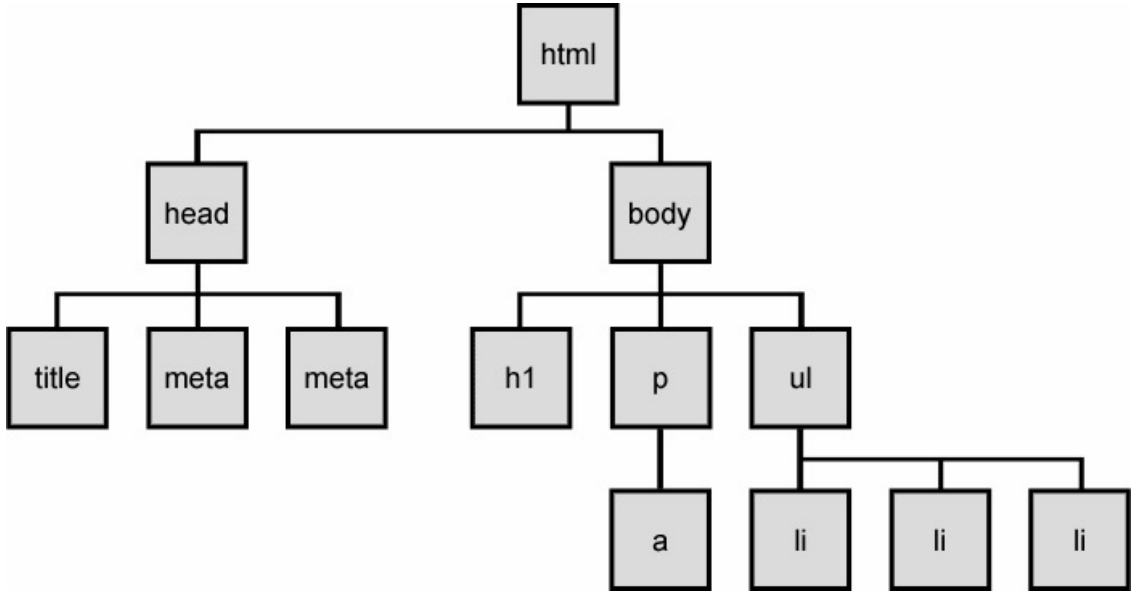


Image Source

HTML In-Lecture Exercise

- Open the following page in Firefox or Chrome

  bit.ly/3kjGmPN

- Right click on “Social Data Science” and click “Inspect”
- What is the hyperlink attribute of the tag? Does it look normal?
  - **2020-2021/STUDYBOARD_2162**
- What are the two tags enclosing it? What do you think they mean?
  - **<li>** and **<ul>**

Cascading Style Sheets

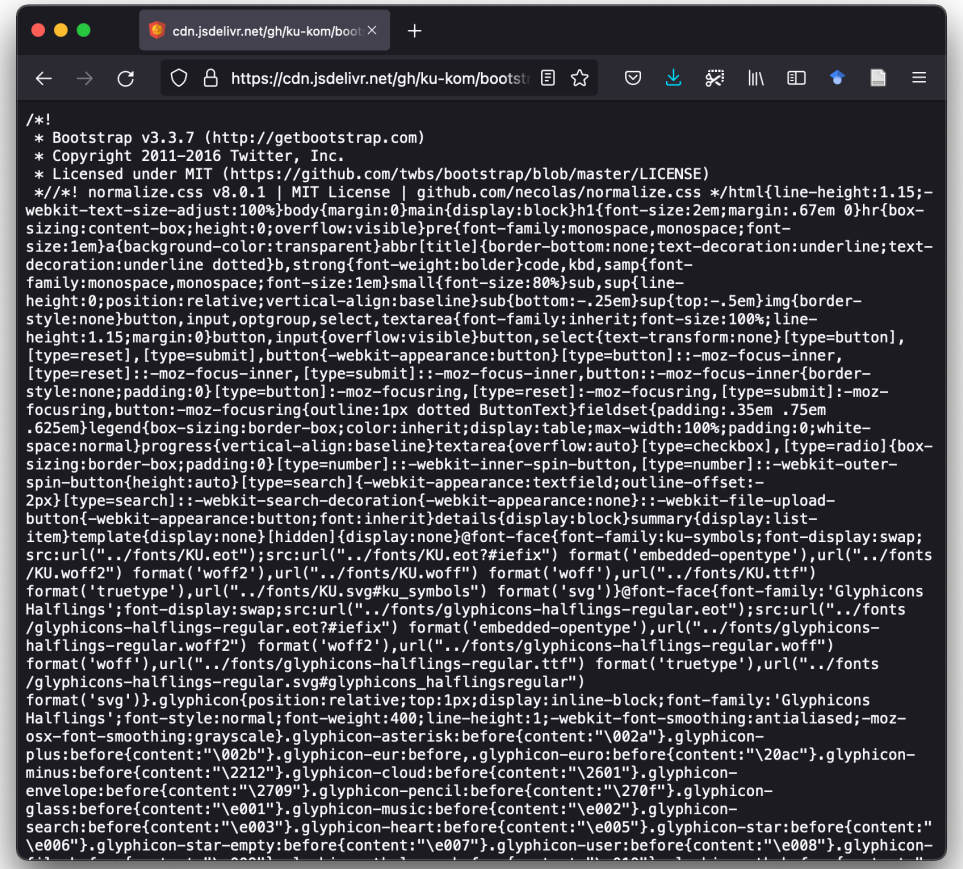
Cascading Style Sheets (CSS)

- CSS is a language to define how documents written in markup are formatted
- It allows separating content from presentation, simplifying the content documents
- Also, benefits page loading speed, adaptation to medium, and uniform presentation

Image Source

CSS in Practice and SDS Applications

- Content document meta-data reference style sheet, which browser loads separately
- Connection between content and style via attributes e.g., `class` or `id`
- Improves content accessibility for data collection via CSS attributes

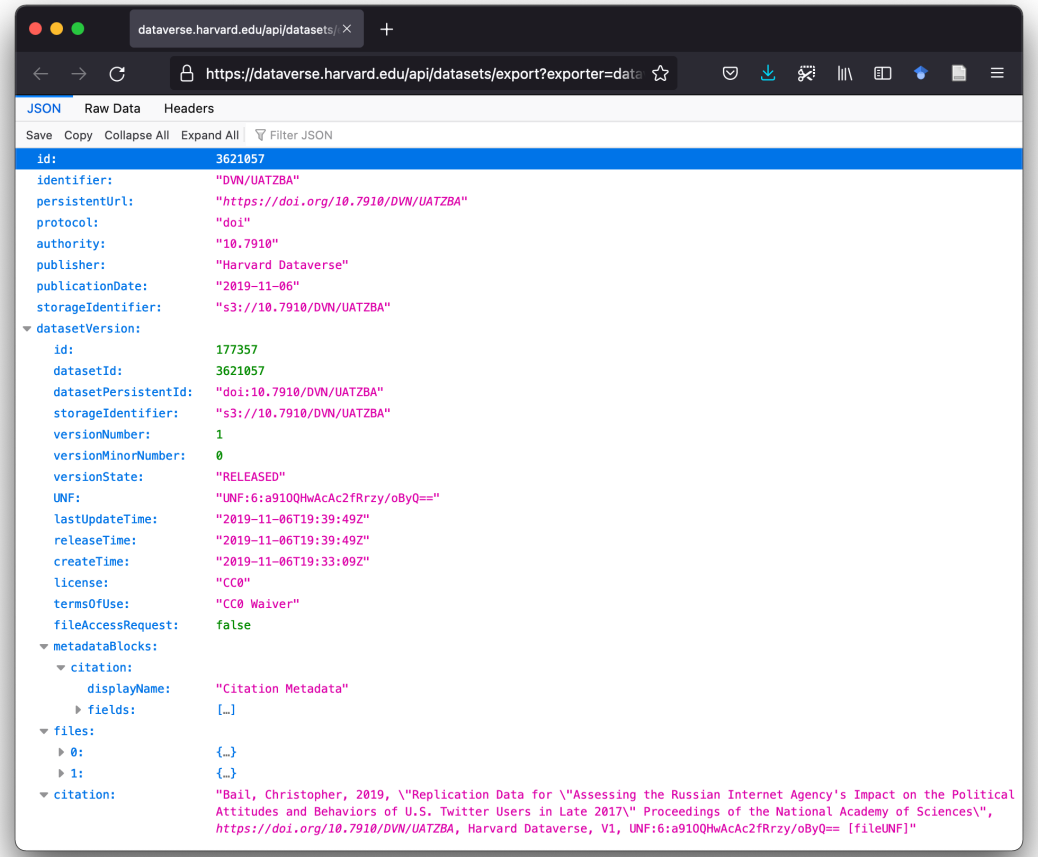


Digital Data Structures

JavaScript Object Notation

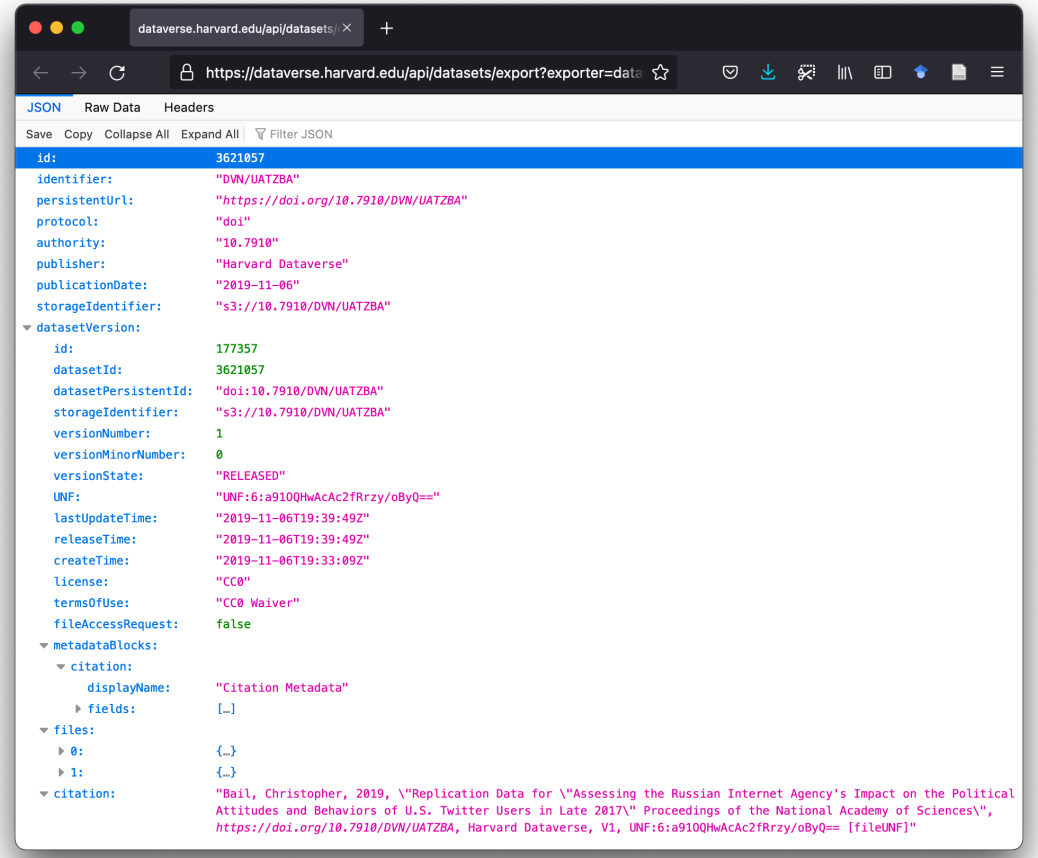
JavaScript Object Notation (JSON)

- JSON is a **non-proprietary**, **language independent** data and file format
- Data stored in **name-value pairs** and arrays
- File extension **.json**



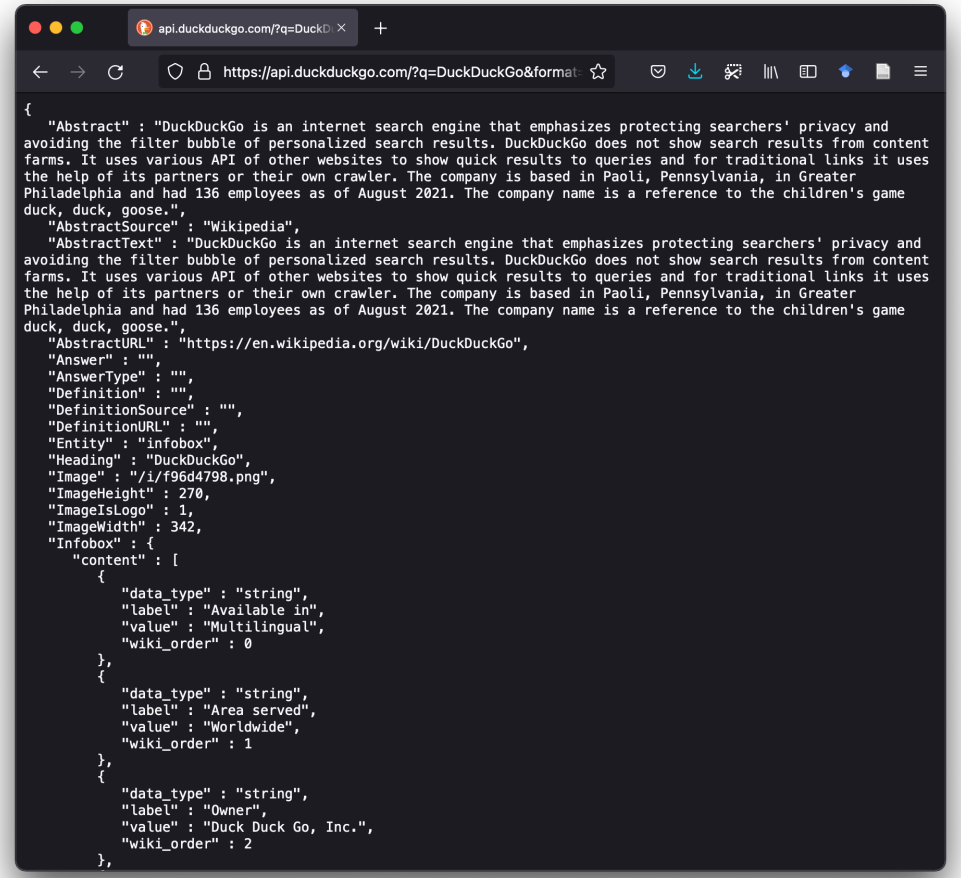
JSON Structure and Syntax

- Hierarchical or tree-like, unordered data structure – akin to Python dictionaries
- Each branch enclosed by curly braces ({ }) with name-value pairs separated by commas
- Values can also be ordered data, in arrays using square brackets ([ ])



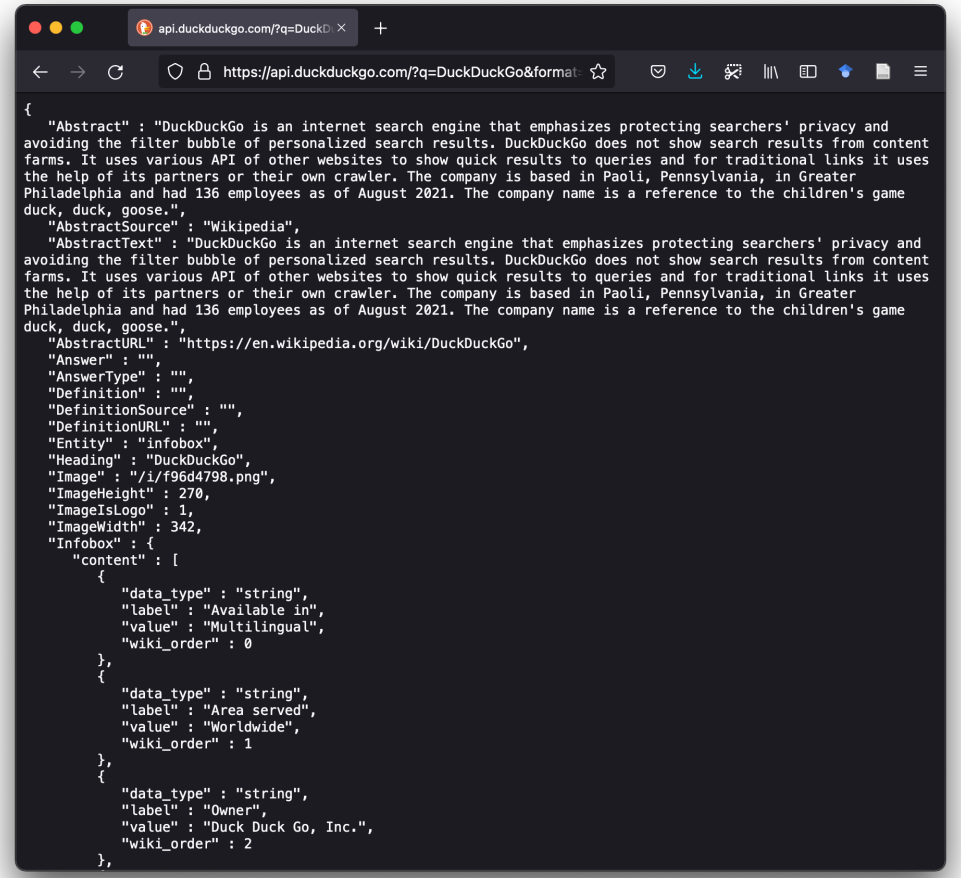
JSON Applications and Sources

- Dynamic and ongoing data collection and distribution, without specified dimensions
- Used e.g., in communication between web applications and servers
- Many APIs provide output in JSON e.g., DuckDuckGo



JSON in Python

- Read in .json as plain text from local file or source
- Use `json` module to decode into Python object (`load`) and encode into json (`dump`)
- Access values using square brackets ([ ]) and names / attributes



Digital Data Structures – Key Take-Homes

1. Non-tabular data formats offer flexibility for data storage and sharing, making them common formats for online data

2. HTML is the common format for web pages, using a hierarchical, tag-based structure to organize and format page content

3. CSS improves formatting work for markup documents, and offers benefits for online data collection

4. JSON is a non-tabular, language independent data format which uses name-value pairs for storage and is commonly used for data storage and sharing online

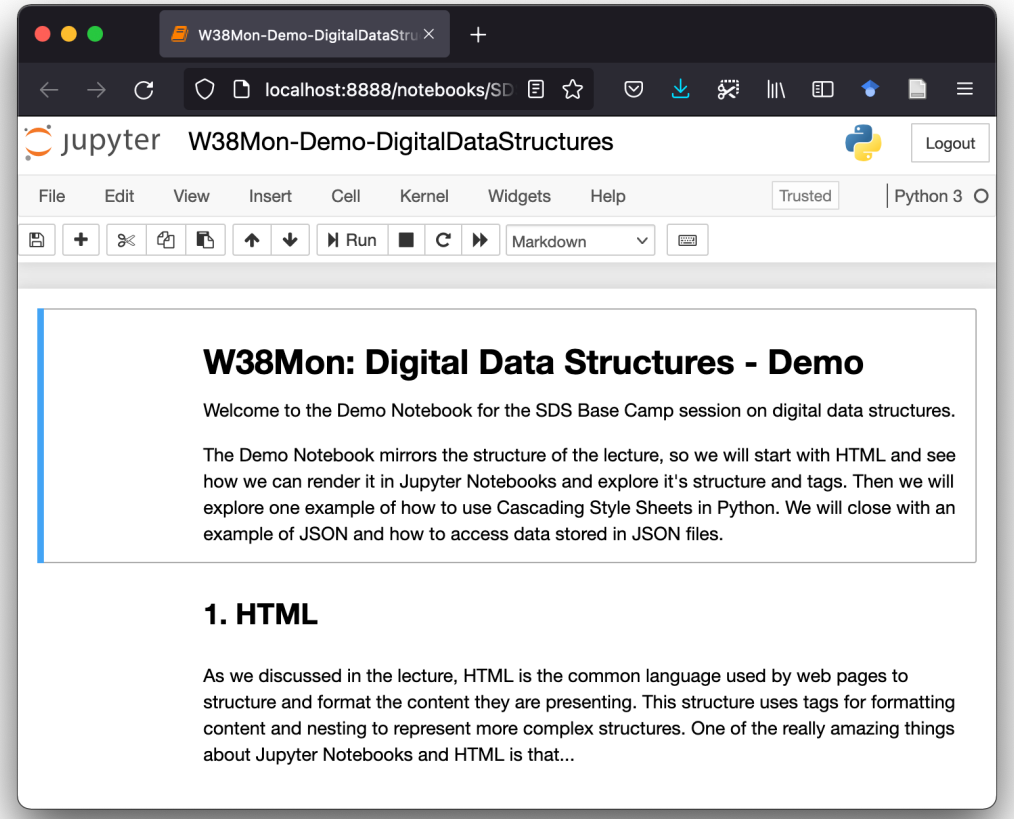
Digital Data Structures

Demo Script

Digital Data Structures – Demo

Let's look at this in practice

- Open "W38Mon-Demo-DigitalDataStructures.ipynb" and follow along



Review Session I Preparation

- Please help us prepare for the first review session, by filling out this short survey on Absalon

bit.ly/3Cm5GuK

"First Review Survey" on Absalon

Exercise Preview

- Exploring HTML in the wild and in Jupyter Notebook
- Exploring CSS in the wild
- Working with JSON in Jupyter Notebook
- Putting it all together
- Daily reflections