Web scraping for research

a brief introduction

Jason T. Kiley, Oklahoma State University

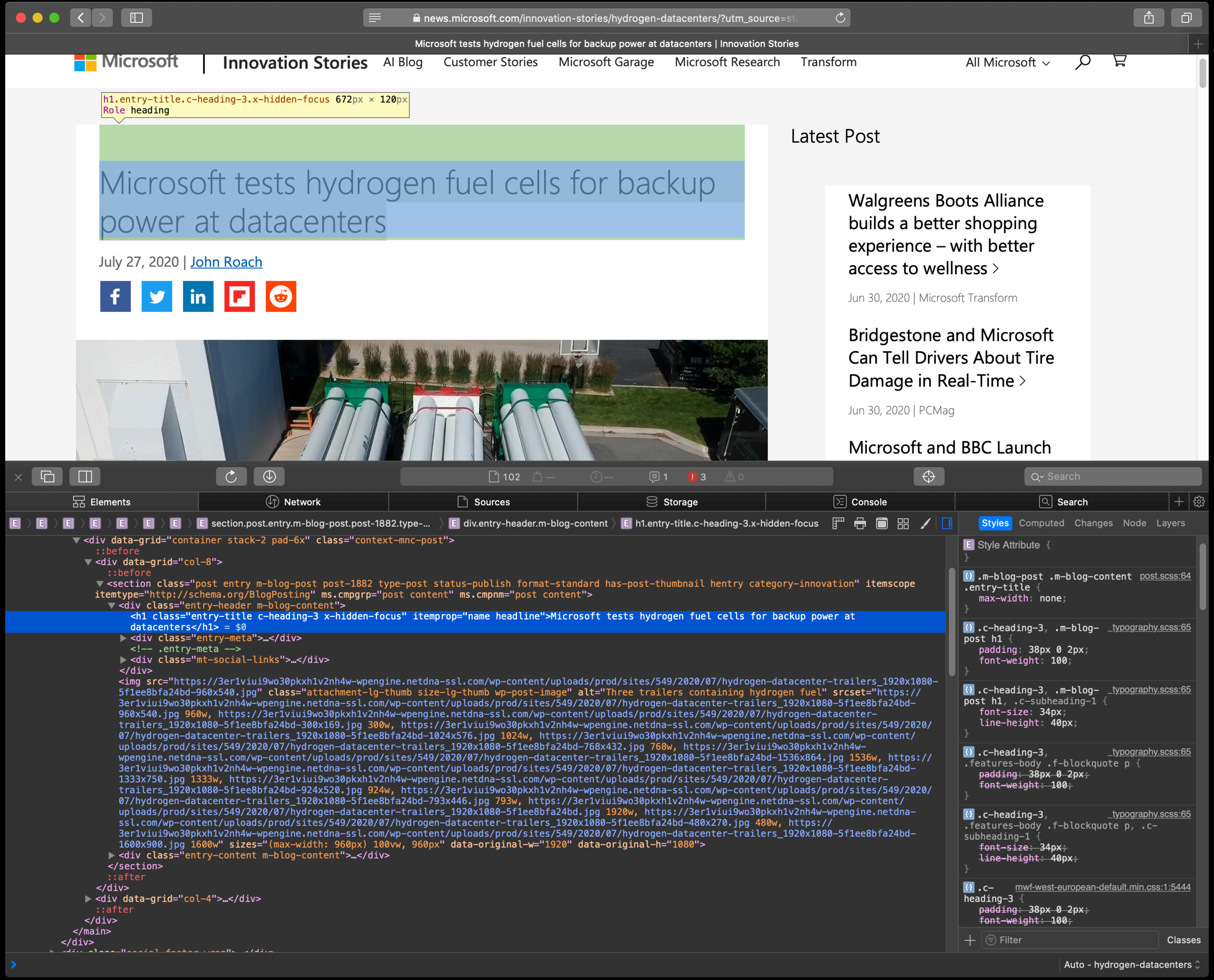
Overview

- What is web scraping?
- When is web scraping most effective?
- What is the process for web scraping data?
- What are the best practices?

Web scraping

defined

- Web scraping is simply extracting usable data from web pages.
- Semi-structured data: there is a structure, but the computer needs us to help see it.



news.microsoft.com/innovation-stories/hydrogen-datacenters/?utm_source=st

Microsoft tests hydrogen fuel cells for backup power at datacenters | Innovation Stories

Microsoft | Innovation Stories   AI Blog   Customer Stories   Microsoft Garage   Microsoft Research   Transform   All Microsoft

```
h1.entry-title.c-heading-3.x-hidden-focus 672px × 120px
Role heading
```

Microsoft tests hydrogen fuel cells for backup power at datacenters

July 27, 2020 | John Roach

Latest Post

Walgreens Boots Alliance builds a better shopping experience – with better access to wellness ›

Jun 30, 2020 | Microsoft Transform

Bridgestone and Microsoft Can Tell Drivers About Tire Damage in Real-Time ›

Jun 30, 2020 | PCMag

Microsoft and BBC Launch

102
1
3
0
Search
Elements
Network
Sources
Storage
Console
Search
section.post.entry.m-blog-post.post-1882.type-…
div.entry-header.m-blog-content
h1.entry-title.c-heading-3.x-hidden-focus
Styles
Computed
Changes
Node
Layers

```
▼<div data-grid="container stack-2 pad-6x" class="context-mnc-post">
   ::before
  ▼<div data-grid="col-8">
     ::before
    ▼<section class="post entry m-blog-post post-1882 type-post status-publish format-standard has-post-thumbnail hentry category-innovation" itemscope
    itemtype="http://schema.org/BlogPosting" ms.cmpgrp="post content" ms.cmpnm="post content">
      ▼<div class="entry-header m-blog-content">
         <h1 class="entry-title c-heading-3 x-hidden-focus" itemprop="name headline">Microsoft tests hydrogen fuel cells for backup power at
         datacenters</h1> = $0
        ▶<div class="entry-meta">…</div>
         <!-- .entry-meta -->
        ▶<div class="mt-social-links">…</div>
       </div>
       <img src="https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/prod/sites/549/2020/07/hydrogen-datacenter-trailers_1920x1080-
       5f1ee8bfa24bd-960x540.jpg" class="attachment-lg-thumb size-lg-thumb wp-post-image" alt="Three trailers containing hydrogen fuel" srcset="https://
       3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/prod/sites/549/2020/07/hydrogen-datacenter-trailers_1920x1080-5f1ee8bfa24bd-
       960x540.jpg 960w, https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/prod/sites/549/2020/07/hydrogen-datacenter-
       trailers_1920x1080-5f1ee8bfa24bd-300x169.jpg 300w, https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/prod/sites/549/2020/
       07/hydrogen-datacenter-trailers_1920x1080-5f1ee8bfa24bd-1024x576.jpg 1024w, https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/
       uploads/prod/sites/549/2020/07/hydrogen-datacenter-trailers_1920x1080-5f1ee8bfa24bd-768x432.jpg 768w, https://3er1viui9wo30pkxh1v2nh4w-
       wpengine.netdna-ssl.com/wp-content/uploads/prod/sites/549/2020/07/hydrogen-datacenter-trailers_1920x1080-5f1ee8bfa24bd-1536x864.jpg 1536w, https://
       3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/prod/sites/549/2020/07/hydrogen-datacenter-trailers_1920x1080-5f1ee8bfa24bd-
       1333x750.jpg 1333w, https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/prod/sites/549/2020/07/hydrogen-datacenter-
       trailers_1920x1080-5f1ee8bfa24bd-924x520.jpg 924w, https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/prod/sites/549/2020/
       07/hydrogen-datacenter-trailers_1920x1080-5f1ee8bfa24bd-793x446.jpg 793w, https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/
       uploads/prod/sites/549/2020/07/hydrogen-datacenter-trailers_1920x1080-5f1ee8bfa24bd.jpg 1920w, https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-
       ssl.com/wp-content/uploads/prod/sites/549/2020/07/hydrogen-datacenter-trailers_1920x1080-5f1ee8bfa24bd-480x270.jpg 480w, https://
       3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/prod/sites/549/2020/07/hydrogen-datacenter-trailers_1920x1080-5f1ee8bfa24bd-
       1600x900.jpg 1600w" sizes="(max-width: 960px) 100vw, 960px" data-original-w="1920" data-original-h="1080">
      ▶<div class="entry-content m-blog-content">…</div>
     </section>
     ::after
   </div>
  ▶<div data-grid="col-4">…</div>
   ::after
 </div>
</main>
</div>
```

```
Style Attribute {
}

.m-blog-post .m-blog-content         post.scss:64
.entry-title {
   max-width: none;
}

.c-heading-3, .m-blog-          _typography.scss:65
post h1 {
   padding: 38px 0 2px;
   font-weight: 100;
}

.c-heading-3, .m-blog-          _typography.scss:65
post h1, .c-subheading-1 {
   font-size: 34px;
   line-height: 40px;
}

.c-heading-3,                   _typography.scss:65
.features-body .f-blockquote p {
   ~~padding: 38px 0 2px;~~
   ~~font-weight: 100;~~
}

.c-heading-3,                   _typography.scss:65
.features-body .f-blockquote p, .c-
subheading-1 {
   ~~font-size: 34px;~~
   ~~line-height: 40px;~~
}

.c-     mwf-west-european-default.min.css:1:5444
heading-3 {
   ~~padding: 38px 0 2px;~~
   ~~font-weight: 100;~~
```

When to scrape

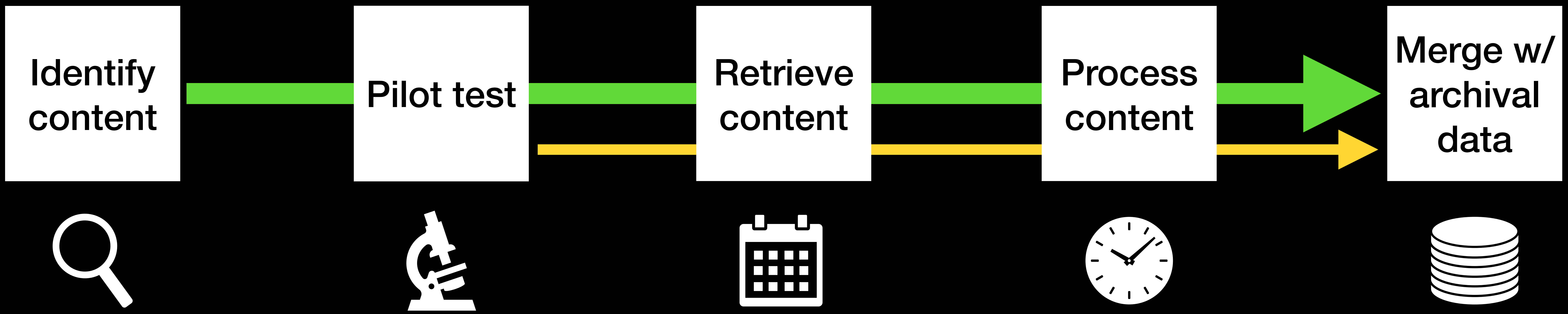
(or not)

- If a site has an application programming interface (API) that works for you, use it.
- If not, consider the number of sources and the amount of data.
- Web scraping has very high fixed costs (per source) and relatively low variable costs.

	One (few) sources	Many sources
Small data	Manual(-ish) gathering	Manual(-ish) gathering
Big data	Web scraping (if no API)	Look for a database or aggregator

Web scraping

the process



Best practices

and hard-won lessons

- Be cool: it’s fairly easy to block web scraping, but most sites only block aggressive scraping (e.g., too fast; too many connections, not targeted). I like to request no more than one page every 10-20 seconds.

- Pilot study: do it. This is a heavy lift in terms of project management, and you’ll save time later by proving that your process works now.

- Retrieve, then process: if you download all of the pages, you can simply reprocess them if (i.e. when) you need to change something.

- Filter first: many sites have pages with links that lead to the full pages that you want. Use the link page data to filter down to only what you need.

See you in the Q&A!