We've recently made an accessibility improvement to the community and therefore posts without any content are no longer allowed. Please use the spoiler feature or add a short message in the message body in order to submit your weekly challenge. ×

2022-05-26 Updates: Email: If you're not seeing emails be delivered from the Community, please check your spam and mark the Community emails as not junk. Thank you for your patience. ×

SIGN IN

Free Trial

Weekly Challenge

Solve the challenge, share your solution and summit the ranks of our Community!

Also available in | Français | Português | Español | ???

IDEAS WANTED

We're actively looking for ideas on how to improve Weekly Challenges and would love to hear what you think!

SUBMIT FEEDBACK ▸

‹ Weekly Challenge

Challenge #40: Parsing a HTML File

GeneR
Alteryx Alumni (Retired)

Happy Monday... oh wait it's Tuesday already. Sorry for the delay if you are an international Alteryx community member, yesterday the USA and Canada celebrated Labor Day in honor of working people.

Hopefully everyone had fun debugging the Macro last week, the link to the solution for that challenge (#39) is HERE. For this week we look at what needs to be done to process raw HTML data after using the download tool to scrape the web.

One of the features of the Alteryx download tool is that it can pull down the raw HTML code from a web page. This practice sometimes referred to as web scraping is useful when there is embedded data in the page you want to access from Alteryx. The challenge is that the raw HTML needs to parsed to prepare the data for use.

Use case: 5280 Magazine in Denver published a list of the best doctors in the Denver metro area, you need to download that list in database form. (Note the Raw HTML has been provided in the workflow)

Objective: Parse the HTML into a database format containing fields for the ID, Physician, Address, City and Practice

Good luck, I hope you are having fun with these challenges and expanding your knowledge of Alteryx. Thanks to all that participate and have provided feedback.

- **challenge_40_start_file.yxmd**
- **challenge_40_solution.yxmd**
- **downloadeddata.yxdb**

Data Preparation | Intermediate | Join | Preparation | Transform

Share | 8 LIKES | Reply

MattD
Alteryx Community Team

Here's a solution:

▷ Spoiler

Share | 0 LIKES | Reply

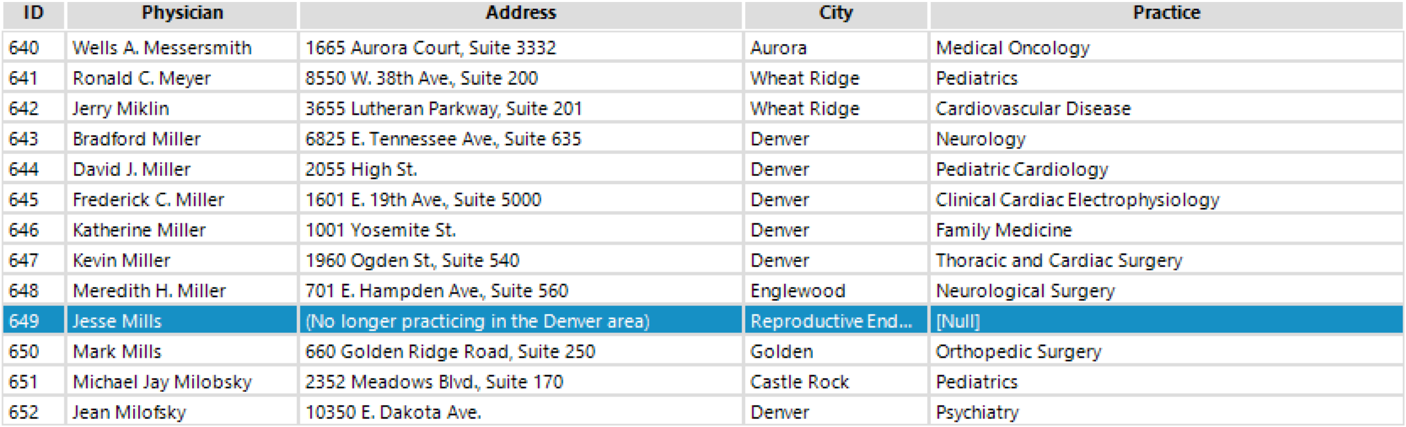
---

brianprestidge
8 - Asteroid

My Solution (I Reg-ex'd the **** out of it!) :-)

FYI - ID 649 was wrong in the provided output solution as the Practice was in the City field:

PS. Loving These Challenges - Keep Em Coming!



ID	Physician	Address	City	Practice
640	Wells A. Messersmith	1665 Aurora Court, Suite 3332	Aurora	Medical Oncology
641	Ronald C. Meyer	8550 W. 38th Ave., Suite 200	Wheat Ridge	Pediatrics
642	Jerry Miklin	3655 Lutheran Parkway, Suite 201	Wheat Ridge	Cardiovascular Disease
643	Bradford Miller	6825 E. Tennessee Ave., Suite 635	Denver	Neurology
644	David J. Miller	2055 High St.	Denver	Pediatric Cardiology
645	Frederick C. Miller	1601 E. 19th Ave., Suite 5000	Denver	Clinical Cardiac Electrophysiology
646	Katherine Miller	1001 Yosemite St.	Denver	Family Medicine
647	Kevin Miller	1960 Ogden St., Suite 540	Denver	Thoracic and Cardiac Surgery
648	Meredith H. Miller	701 E. Hampden Ave., Suite 560	Englewood	Neurological Surgery
649	Jesse Mills	(No longer practicing in the Denver area)	Reproductive End...	[Null]
650	Mark Mills	660 Golden Ridge Road, Suite 250	Golden	Orthopedic Surgery
651	Michael Jay Milobsky	2352 Meadows Blvd., Suite 170	Castle Rock	Pediatrics
652	Jean Milofsky	10350 E. Dakota Ave.	Denver	Psychiatry

▷ Spoiler

Share 1 LIKE Reply

brianprestidge
8 - Asteroid

Hello My Alteryx Friends....

Is this not the same challenge as Week 40 or am i missing something?

Week 40: http://community.alteryx.com/t5/Alteryx-Knowledge-Base/Weekly-Exercise-40-Data-Prep-HTML-Parsing-Dr-...

Share 1 LIKE Reply

GeneR
Alteryx Alumni (Retired)

@brianprestidge you're not missing anything, I must have liked that one so much I posted it twice. I will make sure I have something original for next Monday. Thanks for playing along and keeping us honest!

Share 1 LIKE Reply

brianprestidge
8 - Asteroid

Haha - My pleasure!

I agree, it was a good one so why not do it again!! :-)

Share 1 LIKE Reply

TaraM
Alteryx

A solution has been posted

▷ Spoiler

Tara McCoy

Share 1 LIKE Reply

Joe_Mako
12 - Quasar

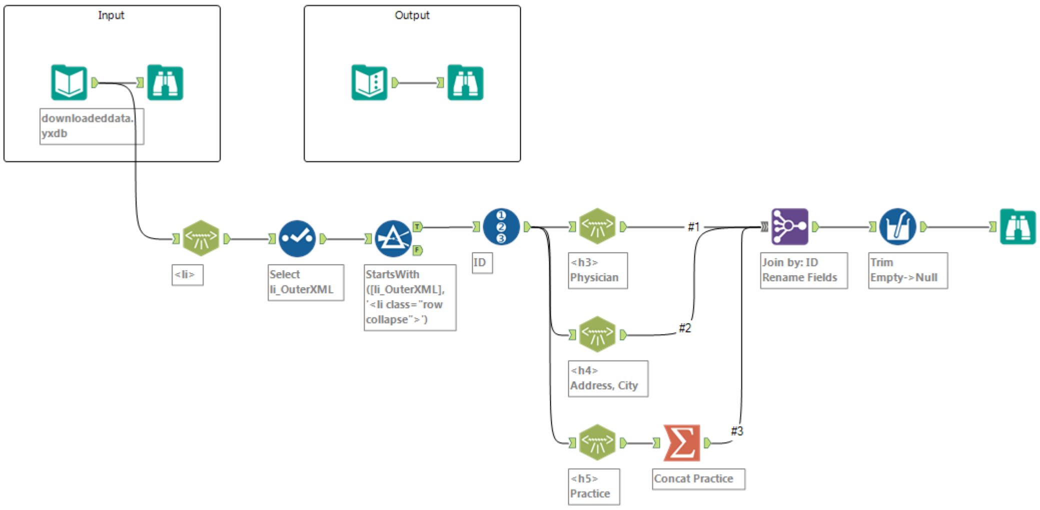
I saw this in another thread, you can find my attached workbook at:
http://community.alteryx.com/t5/Dublin-IRL/Weekly-Exercise-9/gpm-p/36238#M47

Here are three points on the differences between your output and what I came up with:

1. You have an issue with a character encoding in your output
My Results:
493 Yuko Kitahama-D'Ambrosia Denver 4500 E. Ninth Ave., Suite 200 Obstetrics and Gynecology
Your Results:
493 Yuko Kitahama-D&#039;Ambrosia 4500 E. Ninth Ave., Suite 200 Denver Obstetrics and Gynecology

2. For Jesse Mills, the "(..)" text is in the span tag, and in all others the span tag contains the address, but your output has that text in the city field, and then the Practice in the City.
My Results:
649 Jesse Mills (No longer practicing in the Denver area) Reproductive Endocrinology and Infertility [Null]
Your Results:
649 Jesse Mills [Null] (No longer practicing in the Denver area) Reproductive Endocrinology and Infertility

3. 51 physicians have multiple practices, for example, Reginald Bell. Your results only kept the first. I outputted it as a comma separated list in the field.



Share 4 LIKES Reply

GeneR
Alteryx Alumni (Retired)

Nice! @Joe_Mako

Thanks!

Share 0 LIKES Reply

ACE SeanAdams
17 - Castor

I found the same as @Joe_Mako - row 649 in the provided solution has some data corruption.

For @GeneR & @TaraM - for some reason the raw-data for this exercise seems to have dropped of the posting, but it is still available on the link to the Dublin User Group that @Joe_Mako provided below - would you mind adding this to the original challenge posting so that the folk who try this have the data set to work with?

Finally - I felt a little silly when I looked at the posted solution from @TaraM which uses the natural tags to split the data - clearly I did this the hard way.

Have a good weekend all
Sean

challenge_40_SeanSolution.yxmd

Share 0 LIKES Reply

‹ 1 2 3 ... 26 ›

All forum topics ‹ ›

La