DrupalCon

SEATTLE 2019 APRIL 8-12





Migrating Terrible Content to Drupal 8

Intro to Migration Tools

https://events.drupal.org/seattle2019/sessions/migrating-terrible-static-content-drupal-8

Kristian Ducharme CivicActions



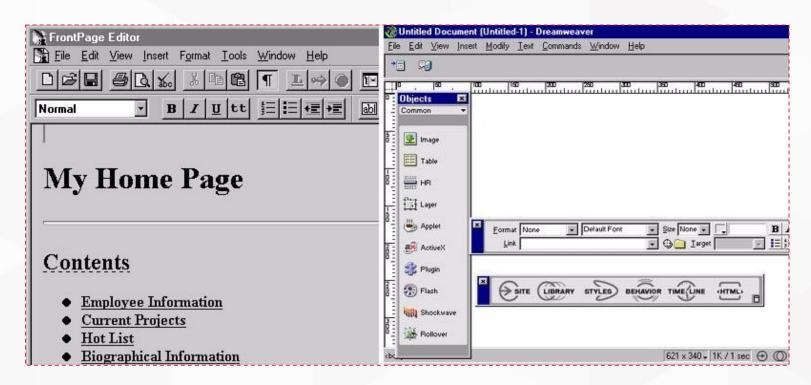
About Me

- ☐ Kristian Ducharme Technical Lead / Product Manager / DevOps Engineer @ CivicActions
- ☐ **Drupal Projects** NSF.gov (current), DEA.gov, DKAN, USDA.gov, Georgia.gov, DigitalDemocracy.org, Whitehouse.gov, City of Los Angeles
- □ Past Presentations DrupalCon Los Angeles 2015, BADCamp 2016
- ☐ What else do I do? Musician, Dad, Electronics DIYer

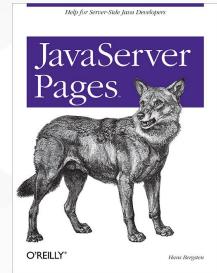


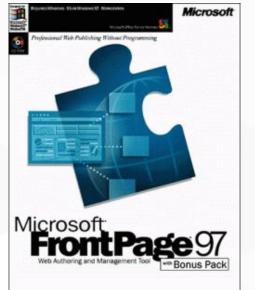
The problem

Almost all websites have "terrible" yet necessary content to migrate.



- A lot has changed since the '90s.
- In most cases, very loose "structure" for static HTML
- Most government sites required to preserve content
- Mobile? Responsive? Accessibility? What's an iPhone?
- Dynamic content was more difficult to make









Difficulties With Static Content Migration

- Source content: Variance in formats/HTML markup/tools used to author
- Varying migration needs: Simple as basic text, as complicated as media w/paragraphs plus file attachments
- **Content buried inside of content:** Tables, deeper links, surrounded by other extraneous information.
- Changing static content before go-live: Needs ability to re-run migrations



Available Drupal Migration Tools

- Core Migrate API: https://www.drupal.org/docs/8/api/migrate-api
- Migrate Plus: https://www.drupal.org/project/migrate_plus (Mike Ryan)
- Migrate Tools: https://www.drupal.org/project/migrate_tools (Mike Ryan)
- Migrate File: https://www.drupal.org/project/migrate_file (Chris Eastwood)
- Migration Tools: https://www.drupal.org/project/migration_tools (CivicActions)
- QueryPath: http://querypath.org



Preparing for Migration (Less "Terrible" Content)

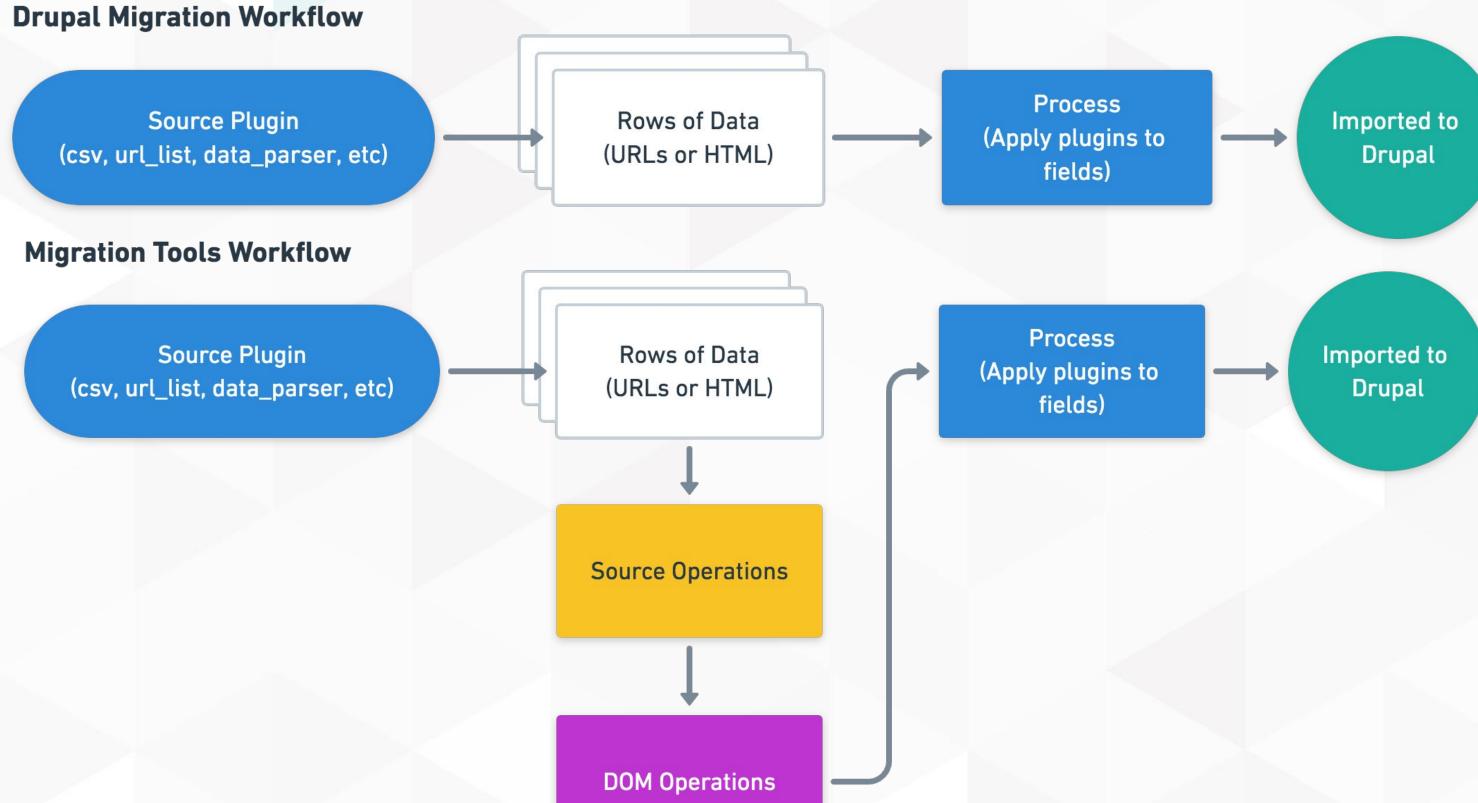
- "Content Cleanup During Migration" Florida DrupalCamp 2019 Steve Wirt
 https://www.fldrupal.camp/sessions/development-performance/content-cleanup-during-migration
- Browser/Spidering Tools Chrome Add-ons: Pesticide, HTML DOM Navigator,
 Site Spider. Screaming Frog
- Auditing Content Spreadsheets for auditing, CSV exporting

	A	В	С	D	E	F	G	Н
1	Source URL	File Type	Content Type	Migrate	Index	Media Import	Page views/mo	Reviewed by
2	http://www.dea.gov/divisions/atl/atl_2015.shtml	html	Press Release	Yes	Yes		1,211	KD
3	https://www.dea.gov/concern/clubdrugs.html	html	Drug Info	Yes	No	Yes	152	JH
4	https://www.dea.gov/concern/cocaine.html	html	Drug Info	Yes	No	Yes	180	JH
5	https://www.dea.gov/concern/concern.htm	htm	Drug Info	Yes	No	Yes	150	JH
6	https://www.dea.gov/concern/ghb.html	html	Drug Info	Yes	No	Yes	192	JH
7	https://www.dea.gov/concern/ketamine.html	html	Drug Info	Yes	No	Yes	242	JH
8	https://www.dea.gov/concern/lsd.html	html	Drug Info	Yes	No	Yes	128	JH
9	https://www.dea.gov/contact-sp.shtml	shtml		No	No		20	AD
10	https://www.dea.gov/contact.shtml	shtml		No	No		696	AD
11	https://www.dea.gov/contactinfo.htm	htm		No	No		150	AD
12	https://www.dea.gov/demand/speakout/index.html	html		No	No		42	AD
13	https://www.dea.gov/directory.htm	htm		No	No		90	AD





Migration Workflow





Configuring Migration Tools

- Migration Tools integrates via PrepareRow, part of "source" configuration.
- Each "Row" can be a URL or HTML data.
- Added to Migration YAML as a "migration_tools" key under "Source" list key.
- Migration YAML
 - Source whether input field is a URL to fetch or HTML content.
 - Source Operations Performed on HTML prior to initializing QueryPath in order specified.
 - Fields Defines jobs for extracting content using Obtainers (May be renamed in future release).
 - o **DOM Operations** Performed on QueryPath object in order specified.



- replaceString
- basicCleanup
- runStringTools
 - fixEncoding
 - convertFatalCharstoASCII
 - convertNonASCIItoASCII
 - stripFunkyChars
 - superTrim
 - stripWindowsCRChars
 - stripCmsLegacyMarkup
 - fixWindowSpecificChars

Source Operations

SourceModifierHTML Class

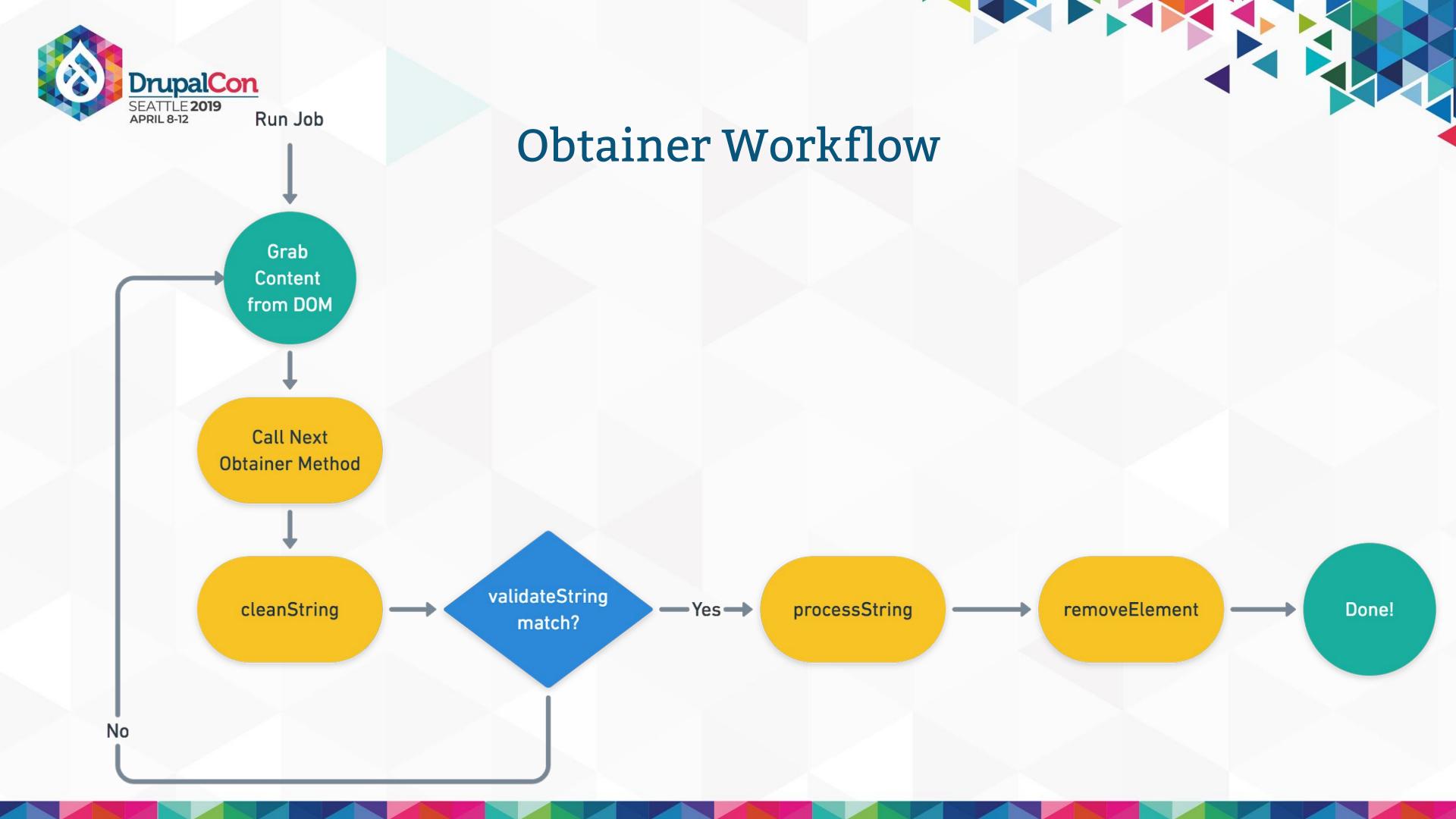
- runStringTools (cont'd)
 - makeWordsFirstCapital
 - reduceDuplicateBr
 - removePhp
 - decodeHtmlEntityNumeric
 - cleanTitle
 - fixHtmlTag
 - fixHeadTag
 - fixBodyTag





Fields Definition

- Name Used by DOM Operations to run this job set
- Obtainer Class to use for obtaining content
- **Jobs** List of jobs to run in order, proceeds until found
 - Job: "addSearch" currently only job type
 - Method: Obtainer method to run
 - **Arguments**: Passed to method





Obtainers

- **ObtainHtml**
- ObtainArray
- ObtainBody
- ObtainCity
- ObtainContentType
- ObtainCountry
- ObtainDate
- ObtainDateSpanish
- ObtainID

- ObtainImage
- ObtainImageFile
- ObtainLink
- ObtainLinkFile
- ObtainLocation
- ObtainState
- ObtainSubTitle
- ObtainTable
- ObtainTitle



DOM Operations

Operation:

- o Get Field Runs jobs defined in the "fields" section
- Modifier Apply a DOM Modifier with arguments

```
# DOM Operations performs the field jobs and applied modifiers in order.
      dom_operations:
          operation: get_field #'get_field' or 'modifier'
          field: title # Field from above to get (run jobs)
          operation: modifier
          modifier: removeSelectorAll
          arguments:
            - '#topbar'
          operation: modifier
          modifier: removeEmptyTables
          operation: modifier
          modifier: removeSelectorAll
          arguments:
            - 'strong'
          # Get the body field after above modifiers have run.
          operation: get_field
          field: body
```



Data Parser Plugin: DOM Parser

- Included with Migration Tools
- What is it? A Migrate Plus module "data parser" plugin (JSON/XML/SOAP)
- What does it do? Allows you to extract URLs from a webpage ("chunking") and process each URL as a "row"
- How do I use it? Combined with Migration Tools, can extract URLs from the DOM

ATLANTA NEWS 2015

Read more division news >>

DEC 02 Nine Methamphetamine Traffickers Sentenced to Federal Prison

DEC 01 Bristol Businessman Sentenced to Serve 24 Months in Prison for Conspiracy to Distribute Synthetic Drugs

NOV 20 Daniel Sands Arrested and Charged with Retaliation against a Federal Witness

NOV 10 Four Sentenced for Involvement in A-PVP Conspiracy

NOV 05 Federal Jury Delivers Guilty Verdict Against Methamphetamine Trafficker

OCT 20 California-Based Heroin Distributor Sentenced to More Than 10 Years in Prison on Drug Trafficking Charges

OCT 20 Joshua Brandon Hinkle Sentenced to 110 Months for A-Pvp Distribution Conspiracy





Source Content:

- HTML Page with list of links to content Determine how to extract links from DOM
- HTML Content Page Determine how to extract elements from a page into Drupal content type fields for migration
- Defining Drupal Content Structure fields (including data only needed for migrating), taxonomies, paragraphs, media, etc.
- Mapping/Extracting content to fields (Migration YAML config)
- Processing leveraging core/contrib migration process plugins





Example: DEA.gov Press Release Archives Listing



```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en-US" lang="en-US">
><head>...</head>
▼<body>
 ▶<div id="wm-ipp" lang="en" style="display: block; direction: ltr;">...</div>
   <!-- BEGIN WAYBACK TOOLBAR INSERT -->
   <div class="wb-autocomplete-suggestions " style="left: 163px; top: 23px; wi</pre>
   <!-- END WAYBACK TOOLBAR INSERT -->
 ▶ <div id="dWideHeaderArea">...</div>
   <a name="SkipToMainContent" id="SkipToMainContent"></a>
 v<div id="dMainContentArea" class="dNineSixty"> == $0
   ▶ <div id="dBreadCrumbNav">...</div>
    <!-- **** Left-hand Column **** -->
   ▼<div id="PLNews-LeftColumn">
      <div class="dColumnHeader">ATLANTA NEWS 2015</div>
     v<div class="PLNews-Article">
      ▼
        ▼
          >...
          ▼
             DEC 02
           ▼>
              <a href="divisions/atl/2015/atl120215.shtml"> Nine Methamphetar
              Prison</a>
             >...
```

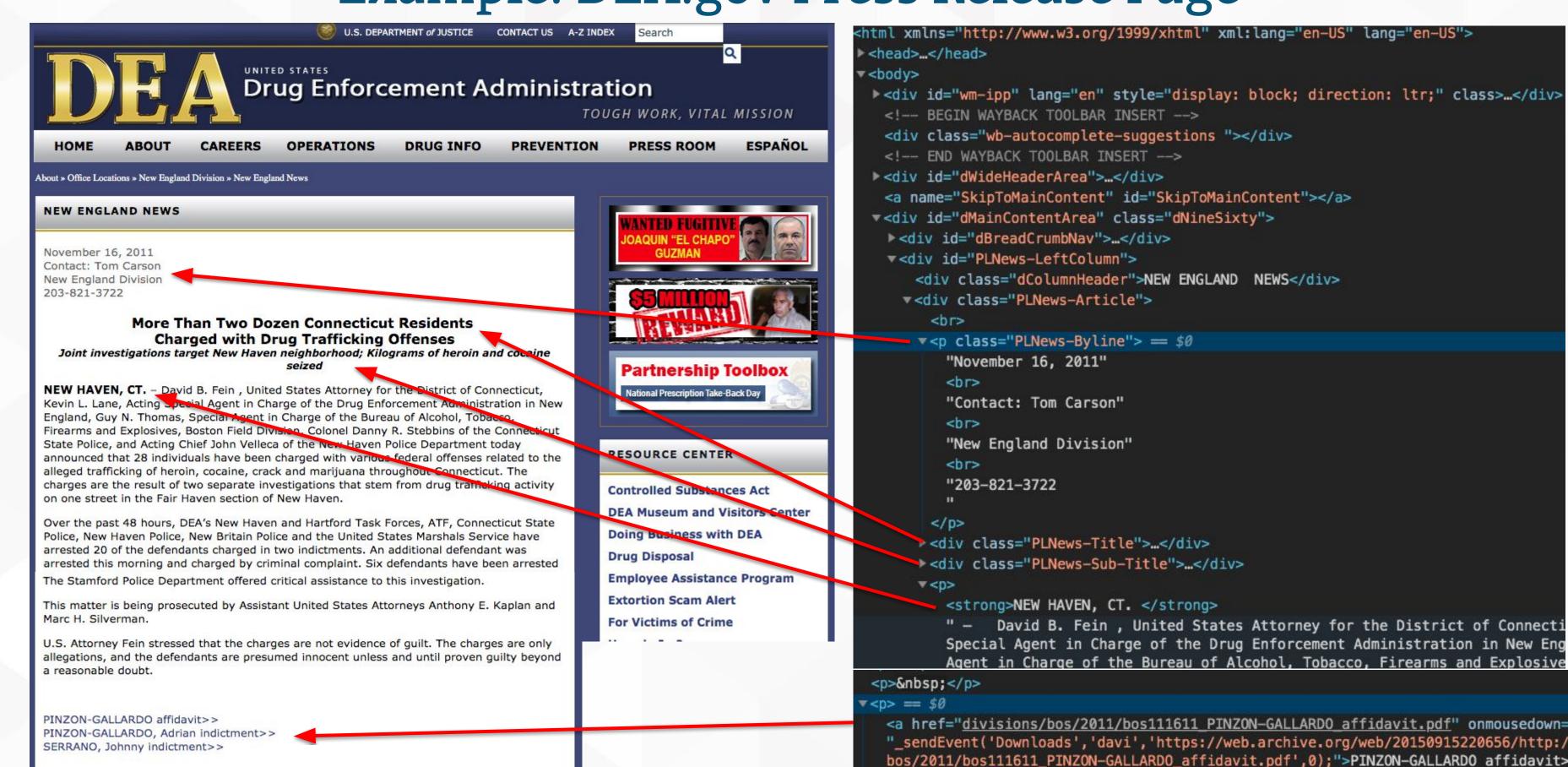
https://web.archive.org/web/20151229193128/http://www.dea.gov/divisions/atl/atl_2015.shtml

Strategy: Press Release Listing Page

- Goal: Capture PR URLS from".PLNews-Article" div area
- Use an **Obtainer** to grab all the URLs from that div:
 ObtainLinkFile, method
 findFileLinksHref
- Base URL or Relative URL links?
 Use a DOM Operation modifier prior to running Obtainer job.

```
source:
plugin: url
data_fetcher_plugin: http
data_parser_plugin: dom
urls:
  - 'https://web.archive.org/web/20150907034317/http://www.dea.gov/divisions/bos/bos_2011.shtml'
ids:
  url:
    type: string
item_selector: url
dom config:
  migration_tools:
      source_operations:
          operation: modifier
          modifier: basicCleanup
      fields:
        url:
          obtainer: ObtainLinkFile
          jobs:
              job: addSearch
              method: findFileLinksHref
              arguments:
                - '.PLNews-Article'
                - [ 'web.archive.org' ]
      dom_operations:
          operation: modifier
          modifier: convertBaseHrefLinks
          operation: get_field
          field: url
```

Example: DEA.gov Press Release Page



https://web.archive.org/web/20150915220656/http://www.dea.gov/divisions/bos/2011/bos111611.shtml



Example: DEA.gov PR Content Type

LABEL	MACHINE NAME	FIELD TYPE			
Attachments	field_pr_attachments	File			
Body	body	Text (formatted, long, with summary)			
Date	field_pr_date	Date			
Division	field_pr_division	Text (plain)			
Imported From URL	field_pr_from_url	Text (plain)			
Phone	field_pr_phone	Text (plain)			
Subtitle	field_pr_subtitle	Text (plain, long)			



Strategy: Press Release Page

- Identify what content needs extraction to fields:
 - Title, Subtitle, Date, Contact, Phone Number,
 Division, Body, PDF Attachments
- Structure of the content:
 - Everything is inside of a "PLNews-Article" div class.
 - Date, Contact, Division, Phone number inside of "PLNews-Byline" div class, separated by

 tags
 - Title is contained in "PLNews-Title" div class,
 Subtitle is in "PLNews-Sub-Title" div class finally an easy one!
 - Body text begins after the Subtitle, contains
 PDF attachment links

Jobs:

- Date From ".PLNews-Byline"?
 How about from URL via regex??
 http://www.dea.gov/divisions/bos/2011/bos111611.shtml = /[a-z]{3}([0-9]+)\.shtml/
- Phone Number Pluck from ".PLNews-Byline", regex: /([0-9]{3}-[0-9]{3}-[0-9]{4})/
- Division from ".PLNews-Byline"? How about from URL via regex??
 http://www.dea.gov/divisions/bos/2011/bos111611.shtml = /divisions\/([a-z]*)\/[0-9]*/
- Title Pluck from ".PLNews-Title"
- Subtitle Pluck from ".PLNews-Sub-Title"
- Body Needs everything above removed before processing so ".PLNews-Article" contains only the body.
- Attachments Pluck files in ".PLNews-Article"



"Subtractive" Content Extraction



```
source:
 plugin: url
 data_fetcher_plugin: http
 data_parser_plugin: dom
 urls:
  - 'https://web.archive.org/web/20150907034317/http://www.dea.gov/divisions/bos/bos_2011.shtml'
 ids:
   url:
     type: string
 item_selector: url
 dom_config:
   migration_tools:
        source_operations:
            operation: modifier
            modifier: basicCleanup
         url:
            obtainer: ObtainLinkFile
            jobs:
                job: addSearch
                method: findFileLinksHref
                arguments:
                  - '.PLNews-Article'
                  - [ 'web.archive.org' ]
        dom_operations:
            operation: modifier
            modifier: convertBaseHrefLinks
            operation: get_field
            field: url
```

```
migration_tools:
                   PR Migration YAML
   source: url
   source type: url
   source_operations:
       operation: modifier
       modifier: basicCleanup
     pdf_files:
       obtainer: ObtainLinkFile
       jobs:
           job: addSearch
           method: pluckFileLinksHref
           arguments:
             - '.PLNews-Article'
            - [ 'pdf' ]
     byline:
       obtainer: ObtainHTML
       jobs:
           job: addSearch
           method: pluckSelector
           arguments:
             - .PLNews-Byline
             - 'innerHTML'
     title:
       obtainer: ObtainTitle
       jobs:
           job: addSearch
           method: pluckSelector
           arguments:
             - .PLNews-Title
```

```
subtitle:
   obtainer: ObtainTitle
    jobs:
        job: addSearch
       method: pluckSelector
       arguments:
          - .PLNews-Sub-Title
  body:
    obtainer: ObtainHTML
    jobs:
        job: addSearch
       method: findSelector
        arguments:
         - .PLNews-Article
         - 'innerHTML'
dom_operations:
   operation: modifier
   modifier: convertBaseHrefLinks
   operation: modifier
   modifier: removeSelectorN
    arguments:
      - '#PLDivision-NewsStoriesTable tr'
      - 1
    operation: get_field
    field: byline
    operation: get_field
    field: title
```

```
operation: get_field
         field: subtitle
         operation: get_field
         field: pdf_files
         operation: get_field
         field: body
process:
 field_pr_date:
     plugin: str_replace
     source: url
     regex: true
     search: '/^.*[a-z]{3}([0-9]+)\.shtml/'
     replace: \1
     plugin: format_date
     from_format: mdy
     to_format: Y-m-d
 field_pr_phone:
     plugin: str replace
     source: byline
     regex: true
     search: '/^.*([0-9]{3}-[0-9]{3}-[0-9]{4}).*/'
     replace: \1
 field_pr_division:
     plugin: str_replace
     source: url
     regex: true
     search: '/^.*([a-z]{3})[0-9]+\.shtml/'
     replace: \1
```

```
title: title
  field_pr_subtitle: subtitle
  body/value: body
  body/format:
    plugin: default_value
    default_value: full_html
  field_pr_from_url: url
  field_pr_attachments:
    plugin: file_import
    source: pdf_files
    destination: 'pdfs/'
  type:
    plugin: default_value
    default_value: press_release
destination:
  plugin: 'entity:node'
migration_dependencies: { }
```





Result: Migrated Press Release Nodes

	TITLE	CONTENT TYPE	AUTHOR	STATUS	UPDATED *	OPERATIONS
0	Prosecution Of Oxycodone Distribution Conspiracy Involving Larry Bushey, Jr., And Others Heads To Conclusion	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🕶
0	Rhode Island Auto Dealer Pleads Guilty To Money Laundering	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🕶
	Central Falls Man Charged In Drug Probe By DEA And Woonsocket Police	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🕶
0	Bridgeport Man Sentenced To More Than Nine Years In Federal Prison For Distributing Cocaine	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🕶
	Two Men Involved In Hartford-Area Narcotics Ring Sentenced To Federal Prison	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🕶
	New Haven Man Sentenced To Three Years In Federal Prison For Distributing Crack	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🔻
	Four Pharmacy Technicians Charged With Stealing And Selling Prescription Drugs	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🔻
	Worcester Man Sentenced To Twenty Years For Selling Cocaine	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🔻
0	Four Sentenced For Distribution Of Oxycodone In Rhode Island And Massachusetts	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🔻
0	Joseph Zapatero Sentenced To Fourteen Years In Prison For Selling Drugs In Windsor And Burlington, Vermont	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🕶
	Civil Penalty Settlement Reached With Baldpate Hospital For Operating An Unregistered Methadone Clinic	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🔻
0	Indictments Charge Eighteen With Oxycodone Trafficking	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🕶
	Former PIN Rx Pharmacist Indicted On Kickback And Tax Charges	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🔻
	TSA And Law Enforcement Officers In Nevadaand Florida Charged In Prescription Drug Ring Narcotics Trafficking Network Results In 20 Arrests	Press Release	Anonymous	Unpublished	04/03/2019 - 16:12	Edit 🕶

Edit Press Release More Than Two Dozen Connecticut Residents Charged With Drug Trafficking Offenses 😭

View

Edit

Delete

Devel

Home » More Than Two Dozen Connecticut Residents Charged With Drug Trafficking Offenses »

History

Title *

More Than Two Dozen Connecticut Residents Charged With Drug Trafficking C

Body (Edit summary)



Result: Migrated Press Release Node Edit

NEW HAVEN, CT. - David B. Fein, United States Attorney for the District of Connecticut, Kevin L. Lane, Acting Special Agent in Charge of the Drug Enforcement Administration in New England, Guy N. Thomas, Special Agent in Charge of the Bureau of Alcohol, Tobacco, Firearms and Explosives, Boston Field Division, Colonel Danny R. Stebbins of the Connecticut State Police, and Acting Chief John Velleca of the New Haven Police Department today announced that 28 individuals have been charged with various federal offenses related to the alleged trafficking of heroin, cocaine, crack and marijuana throughout Connecticut. The charges are the result of two separate investigations that stem from drug trafficking activity on one street in the Fair Haven section of New Haven.

Over the past 48 hours, DEA's New Haven and Hartford Task Forces, ATF, Connecticut State Police, New Haven Police, New Britain Police and the United States Marshals Service have arrested 20 of the defendants charged in two indictments. An additional defendant was arrested this morning and charged by criminal complaint. Six defendants have been arrested previously and are currently in state or federal custody. Two defendants are currently being sought. The indictments and today's arrests represent the culmination of two long-term joint investigations that were funded in significant part by the United States Attorney's Office Organized Crime Drug Enforcement Task Force and supported by the Office's Project Safe Neighborhoods and Anti-Gang programs.

Law enforcement officers executed several search warrants in association with today's arrests and seized at least five firearms, narcotics, cash and other items.

"I want to commend all of the law enforcement officers involved in these important investigations," stated U.S. Attorney Fein. "Their coordinated efforts both today and over the past several months have disrupted what we allege to be two significant narcotics trafficking networks. These cases represent the U.S. Attorney's Office's long-standing, steadfast commitment to working with our federal, state and local law enforcement partners to investigate and prosecute large-scale trafficking of narcotics, and when appropriate, to assisting local law enforcement in their efforts to rid neighborhoods of pervasive drug trafficking in order to improve the quality of life for the residents."

"This investigation is another excellent example of collaborative law enforcement," stated DEA Acting Special Agent in Charge Lane. "By working together, law enforcement agencies were able to dismantle this interstate heroin trafficking group. DEA and our federal, state and local law enforcement counterparts are committed to making our communities safer by removing such drug traffickers from the streets. The arrests in this operation are a positive step in that direction."

"Guns are tools of the trade for drug dealers, as evidenced by today's seizure of firearms," stated ATF Special Agent in Charge Thomas. "ATF will continue to utilize all of its resources to fight violent crime and keep our neighborhoods safe."

"Law Enforcement teamwork is to be credited for the success in this case, which has made a significant impact in the illegal narcotics trade in Connecticut and surrounding states," stated Col. Stebbins of the Connecticut State Police.

"Today's take-down is an example of the difficult and dangerous work the men and women of the New Haven Police Department continue to engage in with our partners in state and federal law enforcement to address organized narcotics dealing and its related gun violence, which affect the city and surrounding region," stated Acting Chief Velleca of the New Haven Police Department. "New Haven residents are encouraged to report any illegal drug activity by calling our tip-line at 203-946-6098. All tips are kept confidential."

ate	
11/16/2011	
one	
03-821-3722	
ivision	
oos	
ubtitle	
oint Investigations Target New Haven Neighborhood; Kilograms Of Heroin And Cocaine S	eized
▼ ATTACHMENTS	
FILE INFORMATION	
→ bos111611_PINZON-GALLARDO-indictment.pdf	
→ ■ bos111611_SERRANO-indictment.pdf	
Add a new file	
Choose Files No file chosen	
Unlimited number of files can be uploaded to this field. 2 MB limit. Allowed types: pdf txt.	
nported From URL	
https://web.archive.org/web/20150907034317/http://www.dea.gov/division	
Published	
Save Preview <u>Delete</u>	

Result: Migrated Press Release Node Edit



Contact Information

Kristian Ducharme

kristian.ducharme@civicactions.com

Drupal.org: kducharm LinkedIn: kristianducharme GitHub: kducharm

CivicActions

http://www.civicactions.com

Thank Yous:

Steve Wirt, Gerardo Gonzales, Beth Jacobson, CivicActions









Friday, April 12, 2019

Mentored Contribution

> 9:00-18:00 Room: 602

First Time Contributor Workshop

> 9:00-12:00 Room: 606

General Contribution

> 9:00-18:00 Room: 6A

#DrupalContributions





What did you think?

Locate this session at the DrupalCon Seattle website:

https://events.drupal.org/seattle2019/sessions/migrating-terrible-static-content-drupal-8

Take the Survey!

https://www.surveymonkey.com/r/DrupalConSeattle

