

Static Search:

An Archivable and Sustainable Search Engine for
the Digital Humanities

Joseph Takeda (@joeytakeda)

(Digital Humanities Innovation Lab at Simon Fraser University)

Martin Holmes

(Humanities Computing and Media Centre at the University of Victoria)

Project Endings

Project Endings

- Static Websites
 - No server side dependencies
 - Strictly XHTML + CSS + Vanilla JS
 - Best chance for preservation 50 years + (we hope)

But...

But...

- Researchers require robust searching mechanisms, including:

But...

- Researchers require robust searching mechanisms, including:
 - Keyword search

But...

- Researchers require robust searching mechanisms, including:
 - Keyword search
 - Exact phrase search

But...

- Researchers require robust searching mechanisms, including:
 - Keyword search
 - Exact phrase search
 - Filtered search (by date, by author, +++)

Why not use...?

Why not use...?

Google Search

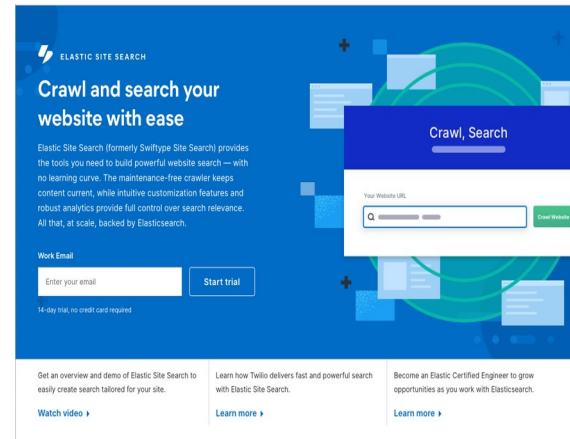
ENHANCED BY Google



Why not use...?

Google Search

ENHANCED BY Google



ELASTIC SITE SEARCH

Crawl and search your website with ease

Elastic Site Search (formerly Swiftype Site Search) provides the tools you need to build powerful website search — with no learning curve. The maintenance-free crawler keeps content current, while intuitive customization features and robust analytics provide full control over search relevance. All that, at scale, backed by Elasticsearch.

Work Email

Enter your email [Start trial](#)

14-day trial, no credit card required

Crawl, Search

Your Website URL [Crawl website](#)

Get an overview and demo of Elastic Site Search to easily create search tailored for your site. [Watch video >](#)

Learn how Twilio delivers fast and powerful search with Elastic Site Search. [Learn more >](#)

Become an Elastic Certified Engineer to grow opportunities as you work with Elasticsearch. [Learn more >](#)

Why not use...?

Google Search

ENHANCED BY Google



HOME DOCS GUIDES DEMO SOURCE

LUNR

SEARCH MADE SIMPLE

GET STARTED

SIMPLE
Designed to be small, yet full featured, Lunr enables you to provide a great search experience without the need for external, server-side, search services.

EXTENSIBLE
Add powerful language processors to give more accurate results to user queries, or tweak the built-in processors to better fit your content.

EVERYWHERE
Lunr has no external dependencies and works in your browser or on the server with node.js

ELASTIC SITE SEARCH

Crawl and search your website with ease

Elastic Site Search (formerly Swiftype Site Search) provides the tools you need to build powerful website search — with no learning curve. The maintenance-free crawler keeps content current, while intuitive customization features and robust analytics provide full control over search relevance. All that, at scale, backed by Elasticsearch.

Crawl, Search

Your Website URL

Q [input] Crawl results

Work Email

Enter your email Start trial

14-day trial, no credit card required

Get an overview and demo of Elastic Site Search to easily create search tailored for your site.
[Watch video >](#)

Learn how Twilio delivers fast and powerful search with Elastic Site Search.
[Learn more >](#)

Become an Elastic Certified Engineer to grow opportunities as you work with Elasticsearch.
[Learn more >](#)

Why not use...?

Why not use...?

- Not particularly reliable

Why not use...?

- Not particularly reliable
- External dependency = technical debt


Why not use...?

- Not particularly reliable
- External dependency = technical debt
- Too many documents to put in one index

So we built our own

So we built our own

- <https://github.com/projectEndings/staticSearch>



projectEndings / staticSearch

Unwatch 3 Unstar 7 Fork 1

Code Issues 19 Pull requests 0 Actions Projects 0 Wiki Security 0 Insights Settings

A codebase to support a pure JSON search engine requiring no backend for any XHTML5 document collection

search javascript offline xslt porter-stemmer xhtml5 Manage topics

455 commits 3 branches 0 packages 3 releases 2 contributors View license

Branch: master New pull request Create new file Upload files Find file Clone or download

This branch is 6 commits behind dev. Pull request Compare

martinholes Add heading and id to get schemaSpec into documentation. Latest commit 76b3ae9 9 days ago

docs	Add heading and id to get schemaSpec into documentation.	9 days ago
emod	Laying groundwork for #35. Adding mayoral shows (and the process by w...	3 months ago
jenkins	Add routine warning message to build log parser.	5 months ago
js	Fix for issue 50.	last month
lib	Switching from Saxon 9 to Saxon 10.	2 months ago
schema	Add heading and id to get schemaSpec into documentation.	9 days ago
test	New intro section for documentation.	9 days ago
utilities	Laying groundwork for #35. Adding mayoral shows (and the process by w...	3 months ago
xsl	Tweak to doc processing to fix html title.	9 days ago
gitignore	Implemented issue #28: an optional setting pointing to a version file...	4 months ago
LICENSE	Initial commit	11 months ago
README.md	Update readme prior to release.	2 months ago
build.xml	Adding documentation to the build and cleaning up the comments	13 days ago
buildSchema.xml	Move schema processing to Saxon 10.	2 months ago
configTest.xml	Implemented item 2 of issue #21: output folder is now configurable in...	4 months ago
configTestWorthReading.xml	Adding a new config for testing #21	3 months ago
config_emod.xml	Laying groundwork for #35. Adding mayoral shows (and the process by w...	3 months ago
config_moeml.xml	Modifying config files to validate against schema after the removal o...	6 months ago
license_BSD.txt	Adding a BSD license	11 months ago
makeRelease.xml	Release build file tested and working.	4 months ago
staticSearch.xpr	Adding more documentation; adding schema build scenario for Oxygen.	4 months ago

README.md

staticSearch

A codebase to support a pure JSON search engine requiring no backend for any XHTML5 document collection

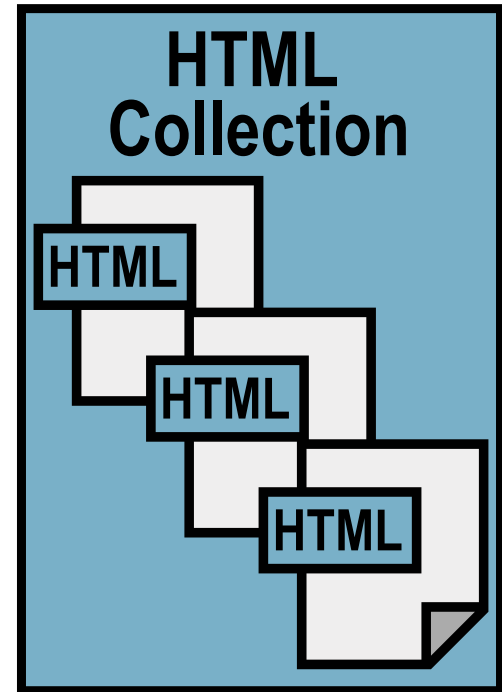
Demonstration applications

- <https://johnkeats.uvic.ca/search.html>
- <https://dvpp.uvic.ca/search.html>

How it works

You have:

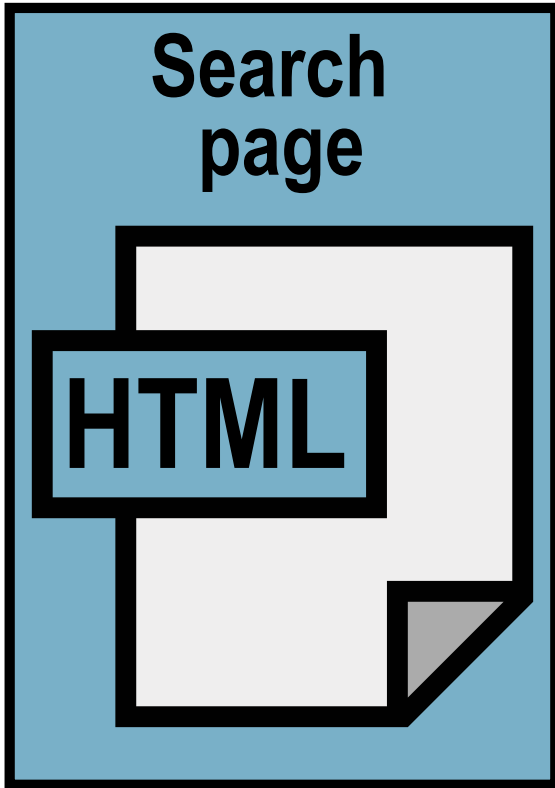
- A collection of HTML files
- All well-formed & valid XHTML



You add:

- An HTML search page containing a special div:

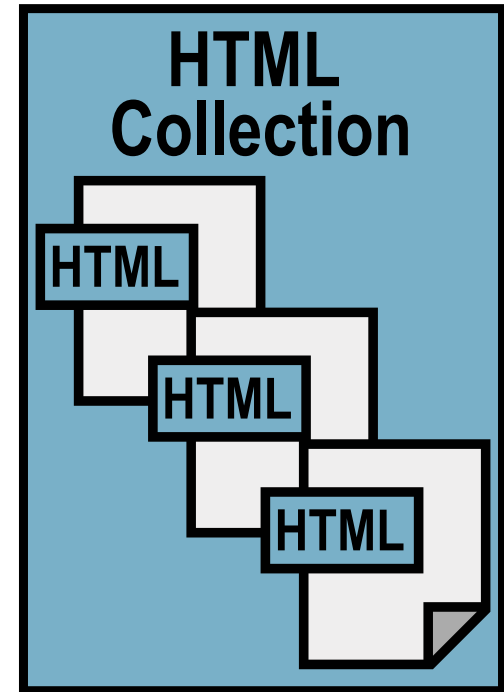
```
<div id="staticSearch"></div>
```



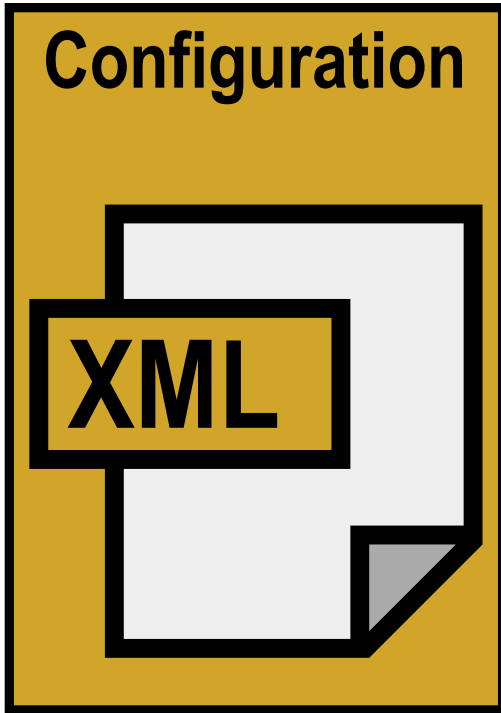
You add:

- Some metadata elements to support search filters:

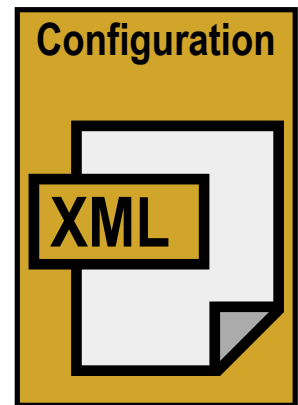
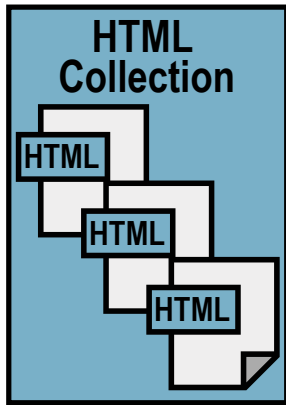
```
<meta name="Document type"  
class="staticSearch.desc"  
content="Poems" />
```

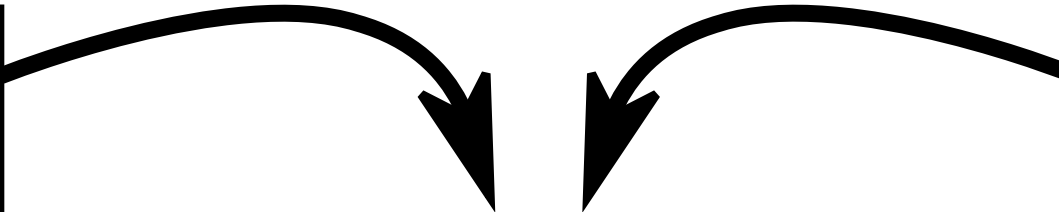
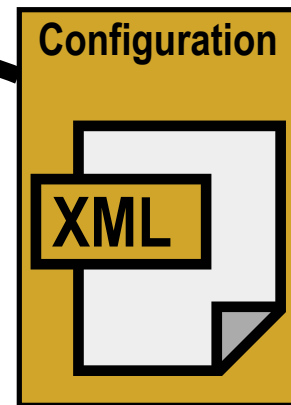
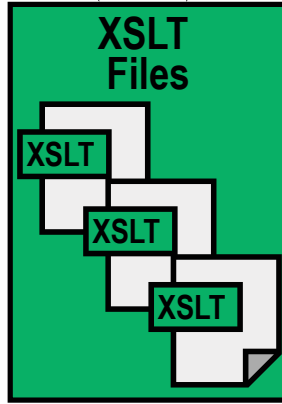
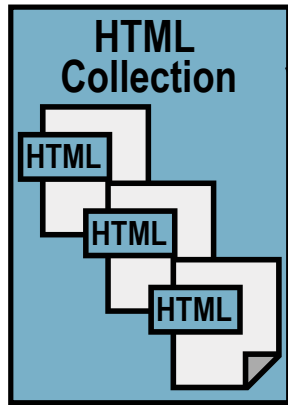


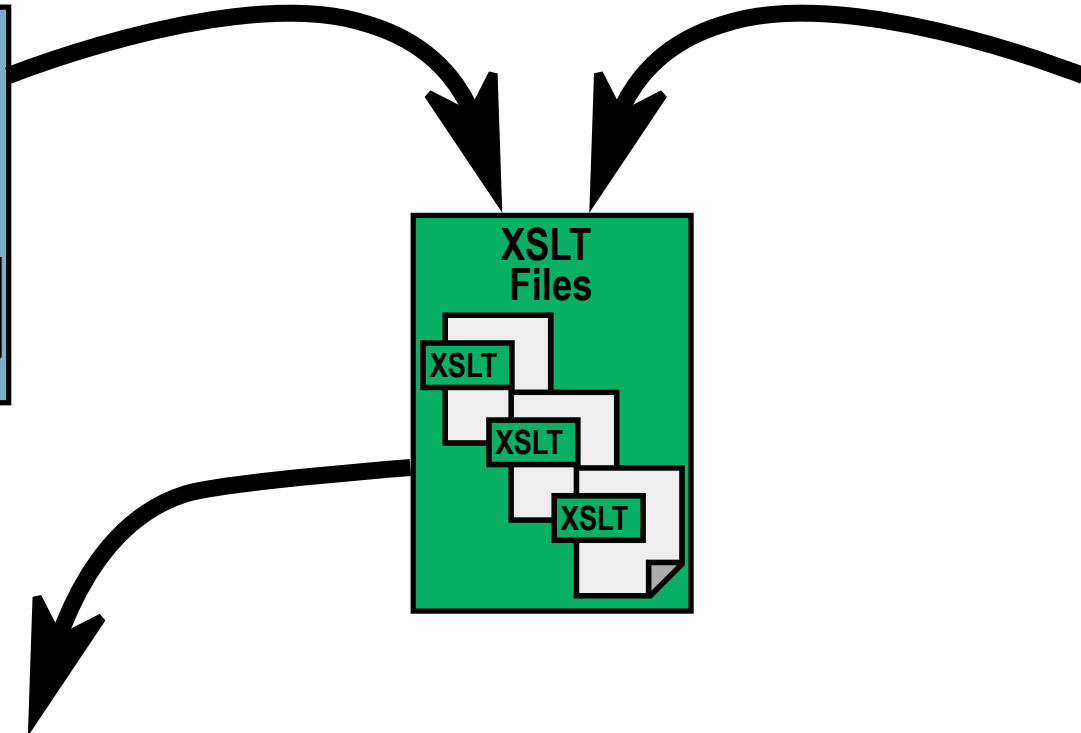
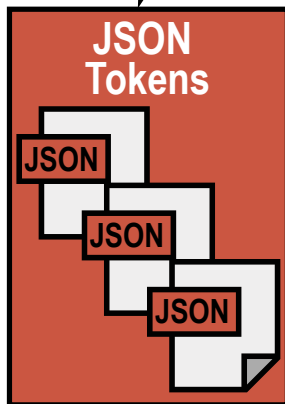
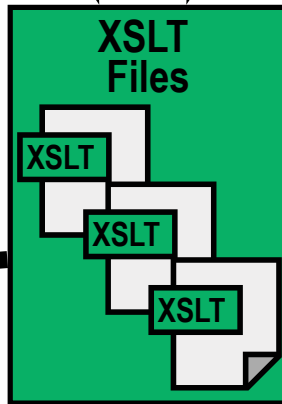
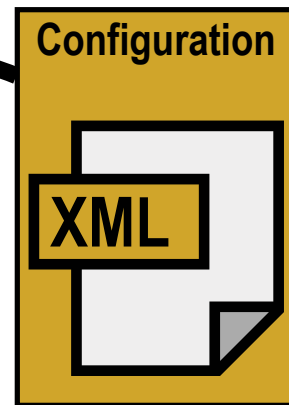
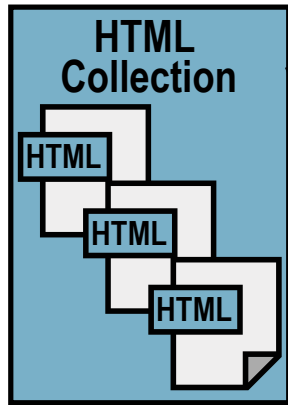
You create:

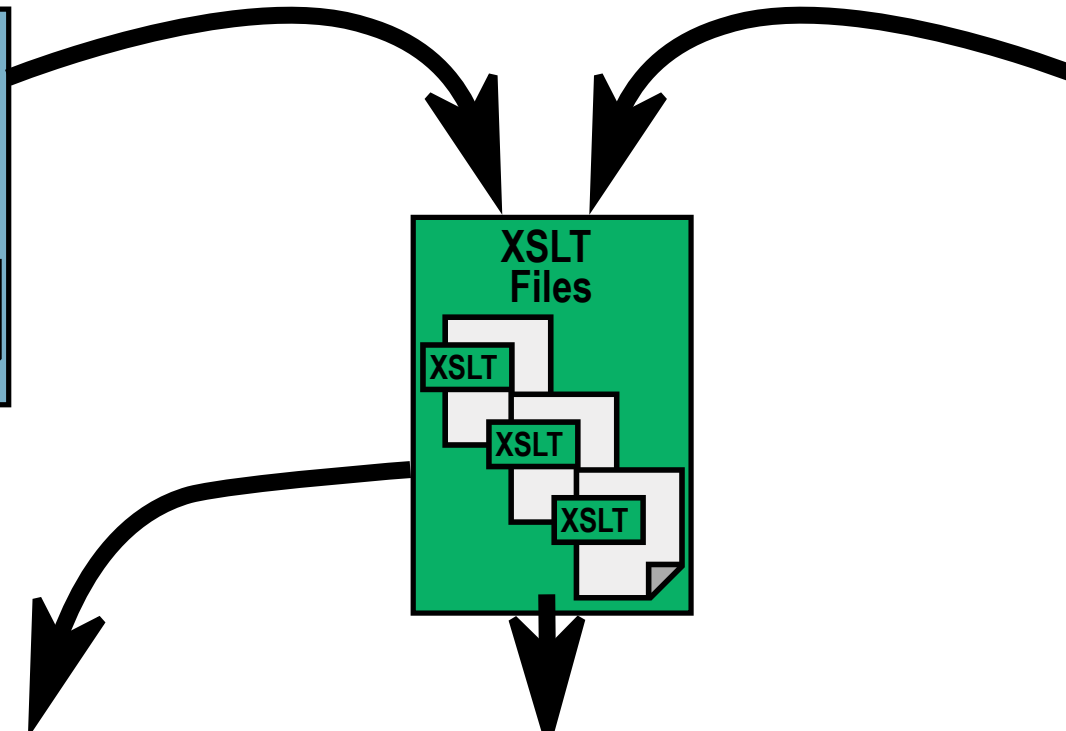
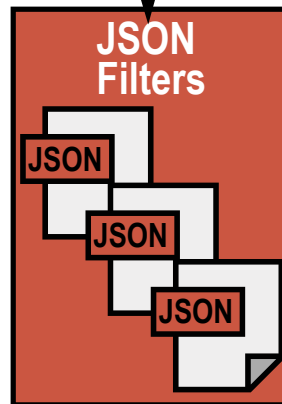
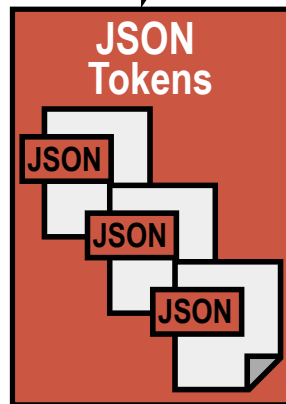
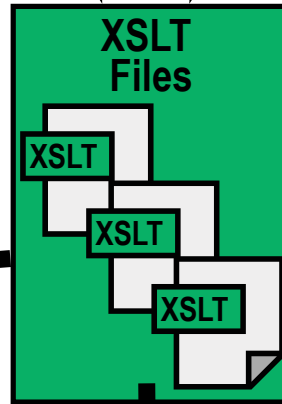
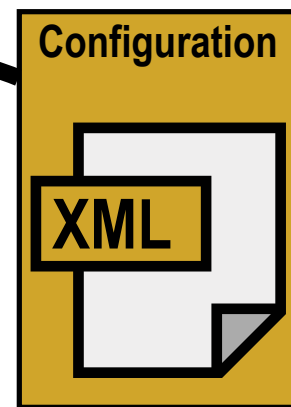
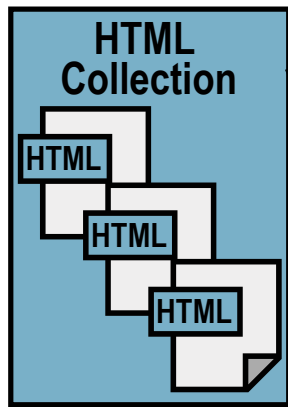


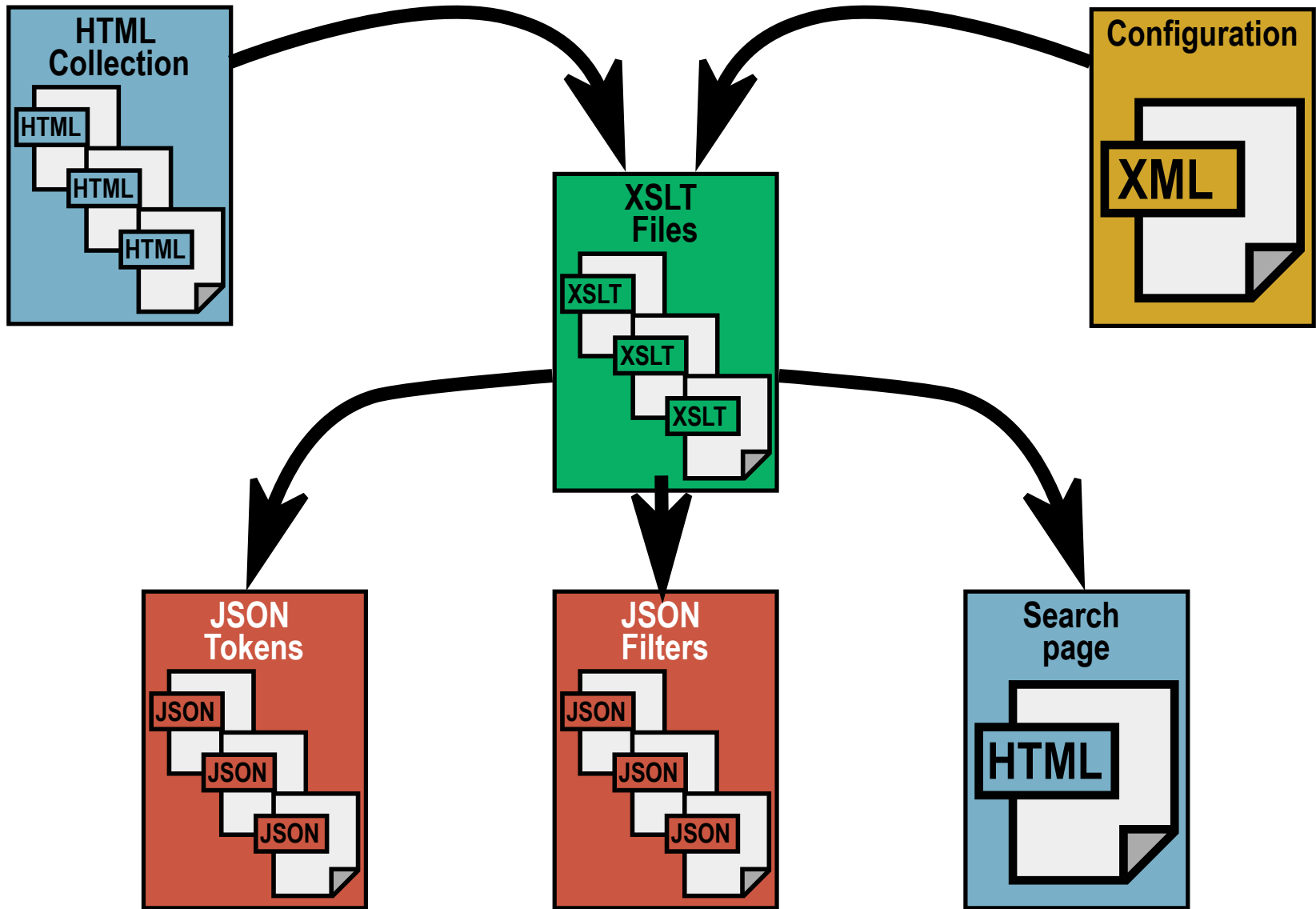
- A configuration file (XML) to specify the features and constraints for your search engine











JSON token file

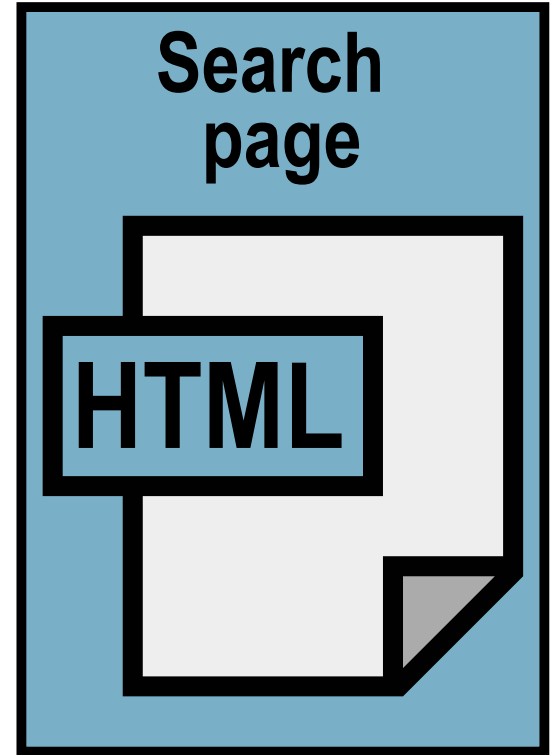
```
{
  "token": "unprofit",
  "instances": [
    {
      "docId": "pom_2025_ithe_old_man_of_hoy",
      "docUri": "poems/goodwords/1870/pom_2025_ithe_old_man_of_hoy.html",
      "score": 1,
      "contexts": [
        {
          "form": "unprofitably",
          "context": "...whole that day was spent <mark>unprofitably</mark>.",
          "weight": 1,
          "pos": 400
        }
      ]
    },
    {
      "docId": "pom_8733_john_and_joan_canto_ii",
      "docUri": "poems/blackwoods/1820/pom_8733_john_and_joan_canto_ii.html",
      "score": 1,
      "contexts": [
        {
          "form": "unprofitable",
          "context": "...too much ap- propriated unto <mark>unprofitable</mark> jocularities and facetiousness. Craving licence,...",
          "weight": 1,
          "pos": 164
        }
      ]
    }
  ]
}
```

Filter JSON

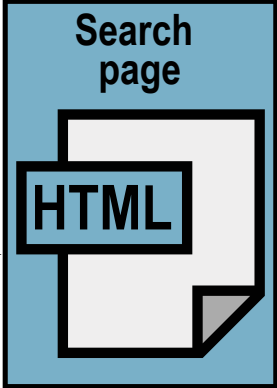
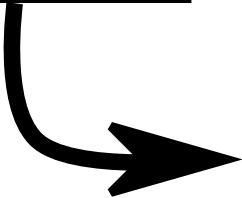
```
{
  "filterId": "ssBool2",
  "filterName": "Unsigned",
  "ssBool2_1": {
    "value": "true",
    "docs": [
      "poems/chambers_series/1867/pom_7594_the_husbands_request.html",
      "poems/alltheyearround/1875/pom_4271_the_hourglass.html",
      "poems/alltheyearround/1879/pom_4507_in_the_conservatory.html",
      "poems/blackwoods/1843/pom_9963_jolly_father_joe.html",
      "poems/blackwoods/1829/pom_10314_the_watchmans_lament.html"
    ]
  }
}
```

On the website...

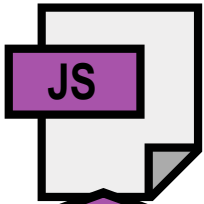
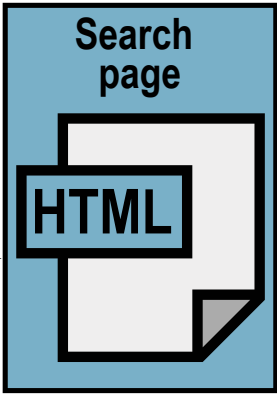
- ...of the thousands of JSON files...
- ...the search page JS retrieves only the ones it needs for the search you do:



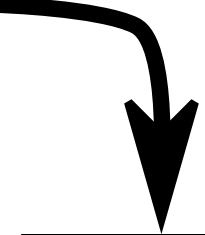
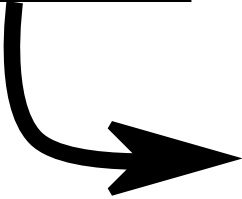
unprofitable

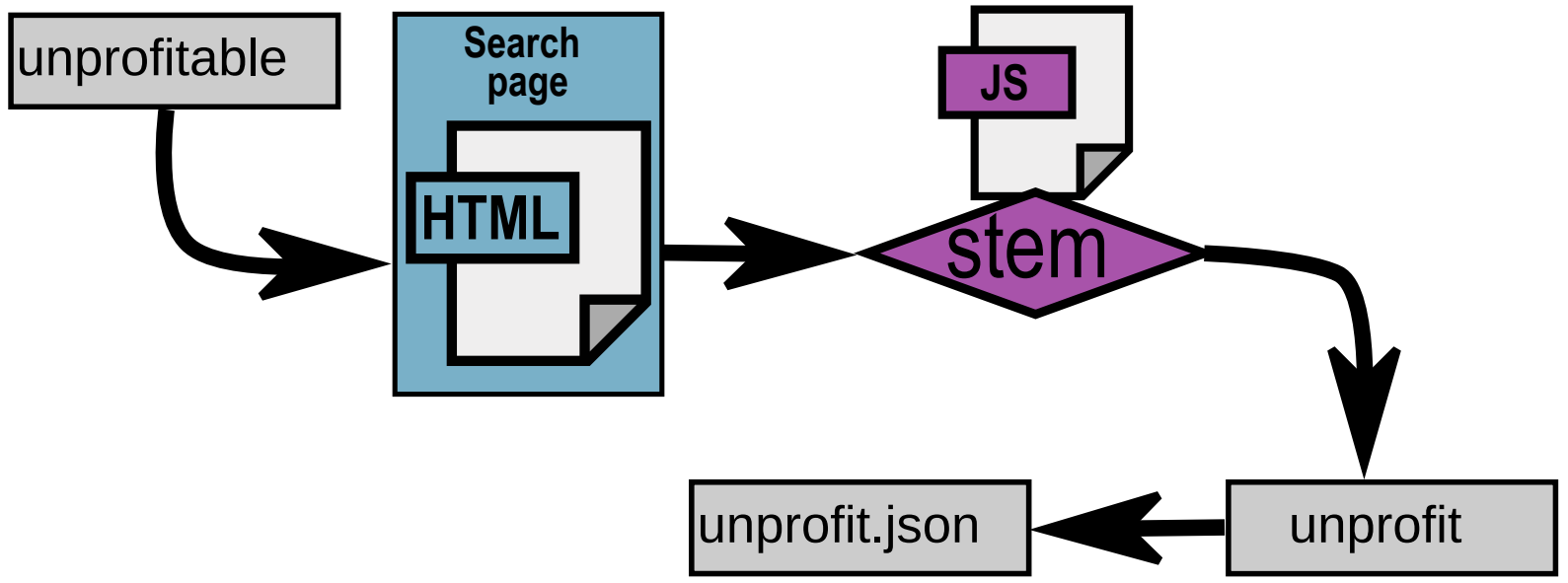


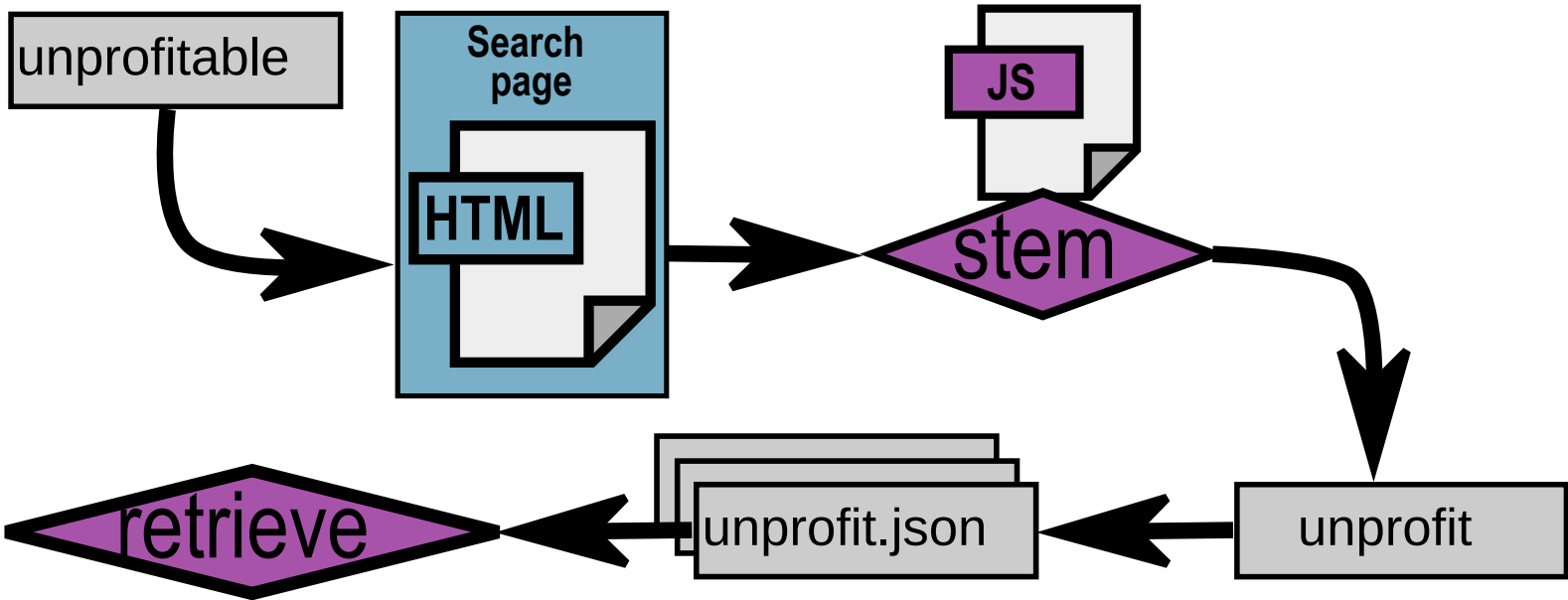
unprofitable

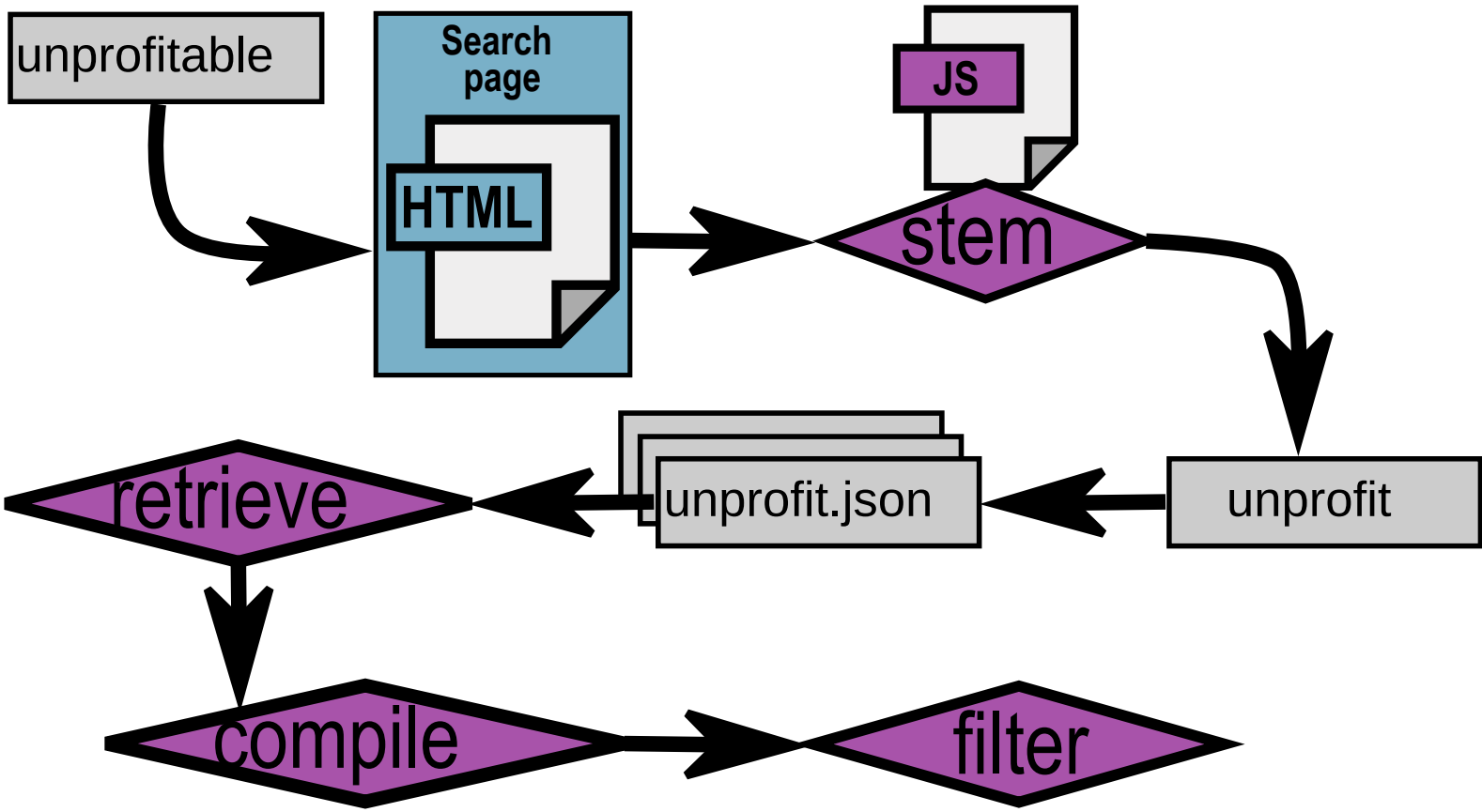


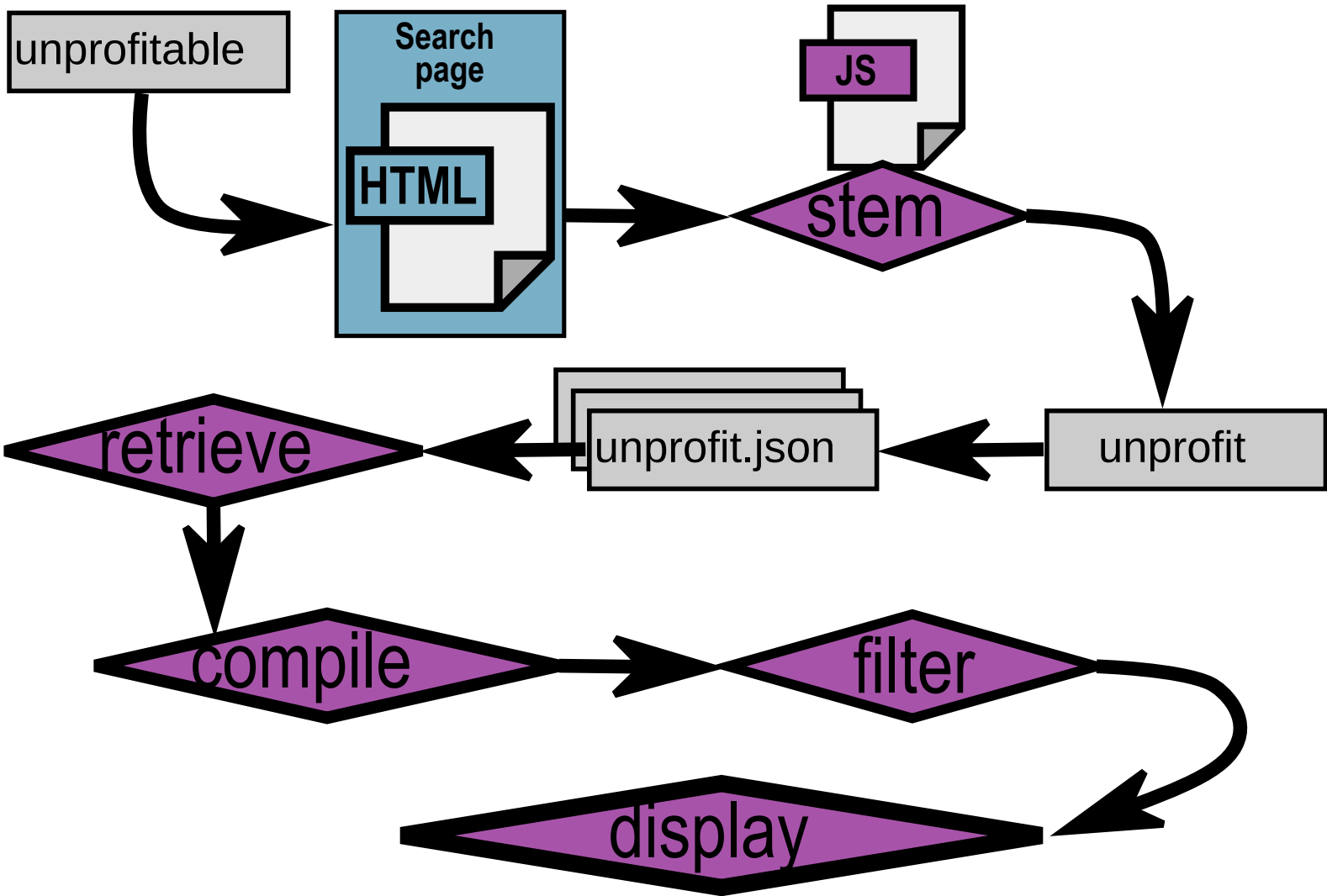
unprofit











Documents found: 2

I.—The Old Man of Hoy Score: 1

- ...whole that day was spent unprofitably.

John and Joan, Canto II Score: 1

- ...too much appropriated unto unprofitable jocularities and facetiousness. Craving licence,...

Next Steps

- Wildcard searches (*, ?, [uv])
- Pluggable stemmers for different languages / dialects

Resources

- Get the code:
<https://github.com/projectEndings/staticSearch>
- Read the documentation:
<https://projectEndings.github.io/staticSearch>

Thanks!

- HCMC, University of Victoria
- DHIL, Simon Fraser University
- Social Sciences and Humanities Research Council
- DHSI, Lindsey Seatter, Arun Jacob