Faceted Search:

User Research Document (URD)

Author(s): Pradyot, Amani

Date: October 19, 2022

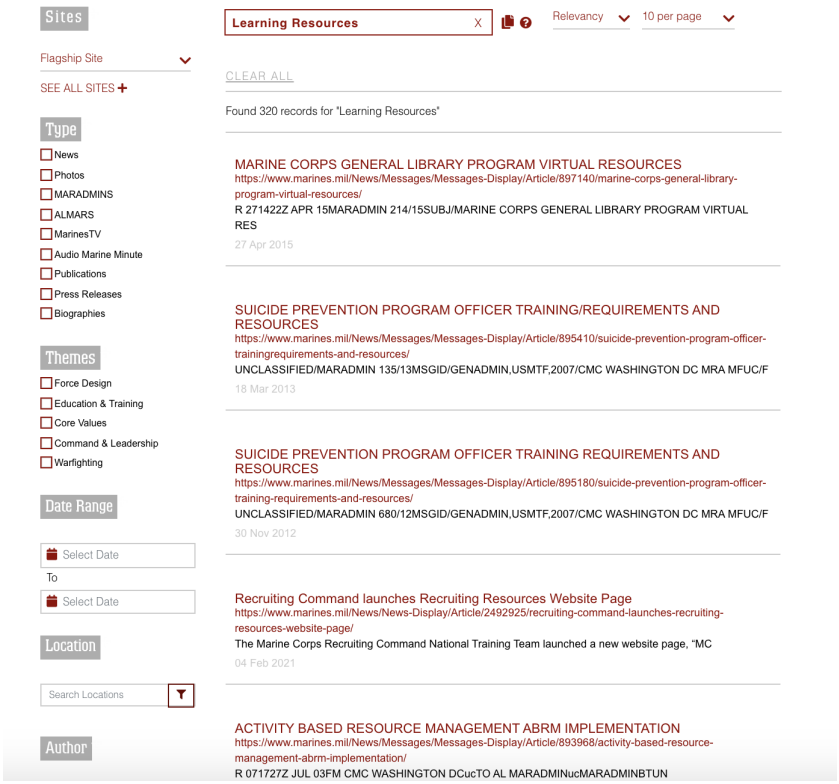
Table of Contents

Research Components

Overview & Objectives: What will this let the customer do? What are our high-level internal goals for doing this project?

Background: Government search inherently covers a wide variety of content - from dense policy and regulation, to pithier public-focused press releases, to time-sensitive information like events and webinars. Because of the breadth of services our partners provide, there are many different types of searchers looking to refine into different content categories on a given website.

Typically, the paradigm searchers are accustomed to (from search experiences like Google, Amazon, etc) when refining content is faceted search - the ability to select the attributes of the content you're looking for to refine the set of results you're presented. An example of a current government site implementing facets is below -



Example search from marines.mil

Problem:

Search.gov was designed to allow the general public to find official government information easily and quickly. We have observed that for searches that return a lot of varied content, research-oriented users aren't able to quickly narrow down into the content most relevant for them, which is causing them to leave our search without finding what they need, or using alternate search methods. How might we improve our search results

experience so that research-oriented users are more successful in finding information quickly?

The only way to segment content in our current search experience is via “Collections”, which silo content into separate verticals (think searching Google News versus. Google Images) rather than allowing users to select the combination of criteria they want to use to refine the search results set. Many of our customers (~40%) have requested this functionality and several large non-customers do not use our service because we do not offer facets.

Proposed Bet:
In order to power the faceted search display, the search experience needs to know what *fields* are available for filtering (ex. Sites, Type, and Themes in the marines.mil example) as well as what *values* those fields contain, so that the user is only presented with options that are applicable to the returned set of documents. This means that for every document in our index, we need to catalog structured, uniform data, in a way that can accommodate the wide variety of agencies and document types we search today. We also need to present it in an intuitive way for the user to be able to refine their search.

Objectives:

- Allow agencies to provide structured information about each document in their index
- Add support for querying across multiple fields with different criteria
- Design an intuitive user interface allowing searchers to refine their search results given the metadata known about a document
- Ensure that site administrators can evaluate the performance of their chosen facets via easy-to-understand analytics views
- Encourage a user-centered approach to facets with our customers, iterating using facet analytics

Terminology Definition: What are key terms referenced in this document?

- **Facet Field:** This is the general field used to store information; ex. Publication Type, Date, Category, Business Unit
- **Facet Label:** This is the heading given to that field; for example, a customer may want to rename the Business Unit field something like “Department”.
- **Facet Values / Facet Options**: These are the values within that field. For example, Publication Type might have Article, Publication, Newsletter, and Press Release as values.

Success metrics: What are the success metrics that indicate you're achieving your internal goals for the project?

- **Engagement:** How many searchers are using the facets in a given session?
- **Improved Usability:** For searchers who use facets, is their click through rate higher and bounce rate lower because they've found the information they need? How is time spent on the search page impacted by using facets?
- **Adoption:** How many agencies are willing to invest the time needed to feature facets on their site?
- **Search Performance:** Are we able to maintain the same standards of query processing time with additional query logic added?

Messaging: What's the product messaging marketing we will use to describe this product to customers, both new and existing?

For common search tasks like finding the right dress (Amazon), a doctor (ZocDoc), or a vacation rental (AirBnB), users are accustomed to filtering across different criteria to help refine the vast amount of results returned to only the options that matter to them. We are now offering faceted search, which surfaces these criteria based off of what exists in your own website. This allows searchers to more easily find the content they're looking for and understand the results that are being presented.

Timeline/release planning: What's the overall schedule you're working towards?

Updated May 17, 2022

Pilot Sites (List to be determined with Admin Team):

- Search.gov
- Invasive Species? (Joyce Bolton)
    - Ensure we have a representative from a site that does not have many resources / support
- USPTO

Investigation for Backend Capabilities (FY 22 Q3, FY 22 Q4)

- **Blockers:**
    - Team Discussion & Consensus on Initial Fields to Support - **Done**
- Dev Investigation Item(s): Explore Options for Supporting Indexing Additional Fields (M1)
    - Infrastructure needed to support expanded index (M2)
    - Where multiple inputs (OpenGraph, Schema.org, Existing Metadata) exist, should we add logic to consolidate into one field, or store separately? (M1)
    - What exists in our customers' websites already? Can we scan a representative set of sites to see percentages that have Schema.org, OG tags, standard metadata, etc? (M1)
        - **Done - SRCH-2937, SRCH-2914**
    - Visualizing indexed content - exposing this to Super Admins (via Kibana?) so that they can guide customers on how to improve their indexed metadata.
    - Determining data type support for "custom" meta fields (S8)
        - **Done**
    - Decisions on data normalization and cleanup of values (C4)
- Dev Investigation Item(s) for Initial Query Capability:
    - Supporting querying across multiple facet fields, and exposing endpoints necessary for frontend development (M3)
    - Capturing analytics on facet selections (S1)
    - Supporting this functionality in the API (S11)

Front End Design (FY22 Q3/Q4)

- **Blockers:**
    - Design resources and bandwidth of USWDS team
        - Meeting notes May 20
- **Partial Blocker:** FE Engineering Resource for consult
- Work with USWDS team to design, usability test, and build a component for faceted navigation that meets all requirements (M5, M6, M7, S2, S5)

- Determine additional support where possible from contractor team resources

Build Backend Support (FY22 Q4 / FY23 Q1)

- **Partial Blocker:** FE Engineering Resource for consult
- Metadata Indexer Development
    - Admin Team to work iteratively with pilot sites to refine their metadata
- Metadata Querying Capability Development
- Pilot sites will inform the initial prototypes

Develop Faceted Search UI (FY22 Q4 / FY23 Q1)

- **Blockers:**
    - SERP Redesign
    - FE Engineering Resource
    - Legacy Front-End Code / Gems
- Dev Investigation Item(s):

Testing and Release (FY23 Q2)

- Usability Testing
- New filter design is the default available for new search configurations, provided data
- Switch current customers to new filter layout as requested
- Write help documentation to add to Search.gov

Personas: Who are the target personas for this product, and which is the key persona?

Searcher: Our end users are the American public accessing these search boxes on the 2000+ federal agency website search boxes we power. They typically enter the search experience from a federal website's header, find the content they need, and return back to that federal website.

- **The researcher (Key Persona):** Whether looking through legal opinions or sifting through patents, these users are browsing through hundreds of technical documents for their purposes. They may want to browse through a couple resources in a given category, just as a place to start. They typically use *broad, unambiguous queries* and narrow down their result set from there.
- **The benefits seeker (Not Intended Audience):** This searcher is coming to a site using Search.gov for quick answers - where can I find a given service. They may not want to spend a lot of time in the search, so features like Best Bets and a stronger ranking algorithm would benefit this subset of searchers much more.

Site Administrator: Our customers are federal agency partners, typically those working in the Web/IT department for an agency that manage their main websites. These site administrators are responsible for both the display and content of their search experience.

- **The search experience optimizer (Key Persona):** These site administrators have both the desire and ability to make customizations to their website to facilitate improved search. They are willing to make development updates to improve SEO and on-site search. They are involved in reviewing their search metrics and always looking for ways to improve the searcher experience.
- **The plug-and-play administrator (Not Intended Audience):** These users want to get a good-enough search running on their site, with little day-to-day involvement. They may not have direct access to the websites themselves, so cannot make website updates. If there are features that can be added without involvement from an engineering team or heavy investment upfront, those would be desired.

User scenarios: These are full stories about how various personas will use the product in context.

Searcher Scenarios

Scenario 1: Searcher enters search, and uses facets to refine query.
Link to PDF Visual

Step 1: Searcher navigates to nasa.gov, and searches for "Planets", in hopes of finding recent news pertaining to planet discovery

a. Searcher is presented with a lot of results. They see that they can refine their query by filtering the organic search result set.
b. Searcher selects the button to refine the results set.

Step 2: Searcher is sent to the website results vertical, where they can discover ways they can refine the results set.

a. Searcher looks through the available facets
b. Searcher decides they only want to view Press Releases from within the last 30 days
c. Searcher selects these two facets, and hits apply

Step 3: Searcher views results

a. With this refined result set, the searcher is able to browse the most recent press releases at their leisure, and may click on one or more results to read.

Scenario 2: Searcher is presented with few or no results that match their selection

1. Searcher selects a facet that only returns 2 results
2. Searcher reviews the snippets for each and determines neither will be helpful
3. Searcher is presented with a prompt at the end of the page 1 results - "Didn't find what you were looking for? Click here to see all results for your query."

4. Searcher selects prompt. Facet applications are removed, and searcher is presented results for the original search term.

Scenario 3: Searcher is sent a link to a query with preselected facets

1. Searcher is reading an article about great day hikes in the California national forests
2. Searcher sees a link to explore hikes in all California forests. Searcher selects link to learn more.
3. Searcher is taken to the National Forest Service search page, with the facets for "Day Hikes" as a Category and several California forests already pre-selected.
4. Searcher can browse through the results for this filtered query, and also explore other activities in these different forests by manipulating the facets.

Site Administrator Scenarios

Scenario 1: Site Administrator sets up facet display

The web manager at www.marines.mil want to set up a faceted search display with Search.gov that mirrors their current experience here: https://www.marines.mil/Home/Search/

1. They know they have data in their content management system that classifies each document by Type, Themes, and Author. They also would like users to be able to filter by date published, as well as the domain the document appears on.
2. They see in their Admin Center a tab to configure facets, and a link to Search.gov documentation to learn how to do so.
3. Reading the documentation, they see that there are a few metadata constructs supported by Search.gov. They decide to implement the below metadata for all documents where they have the information:
    a. Type: <meta custom1="">
    b. Themes: <meta keywords="theme1, theme2">
    c. Author: FirstName LastName
4. They know that the date is already captured correctly from their CMS to Search.gov's index, and the domain is captured in the system, so they now have all the inputs required for facet set up.
5. They finish development, and *request Search.gov reindex their domain* so that we can capture this new information for all documents.
6. They go back into their Admin Center, and set up the mapping and facet labels for the five facets they want to feature:
    a. Type > custom1
    b. Themes > meta keywords
    c. Author > rel="author"
    d. Date > Last Modified Date
    e. Site > domain name
7. They decide to turn on this feature and see how their searchers interact with it.

Scenario 2: Site Administrator views analytics

A few weeks have passed by since marines.mil have added faceted search to their Search.gov experience, and they want to analyze how it's being used.

1. The site administrator logs into their Admin Center, and clicks on the Analytics Tab
2. In the queries tab, they see their top queries for the month. They also see for each query, the percentage of people who engaged with facets for that given query. They're also able to see the CTR in this view.
3. They want to drill a bit further. For the top query, they see they have 20% of users selecting facets. They click to "View Facet Selections" to understand how often, for that query, were specific facet *fields* and *values* used
4. They also want to see the aggregate view. They navigate to the new Analytics tab, Facets. This shows a few important metrics:
    a. Overall percentage of queries that use facets
    b. Click thru rate of searches w/ facet usage vs. those without
    c. Top facet fields used
    d. Top facet values used

Scenario 3: Site Administrator wants to disable facet display

The site administrator, looking at the analytics, decided that the current display is too overwhelming to users. They are going to go back to the drawing board and switch up their document metadata, but in the meantime, want to turn off facets entirely.

1. Site admin logs into the Admin Center
2. Site admin navigates to Display > Facets, and turns off all facet fields.

Scenario 4: Site Administrator adds additional facets and changes display names

The site administrator has worked with their team to streamline the facets they offer, and update the metadata to reflect that.

1. Site admin requests a reindex from the Search.gov team
2. Site admin goes into the Admin center, and adds three facets, disabling the rest.

MoSCoW Table: Overview of all requirements

Must	Should	Could	Won't
M1. We must have the ability to ingest structured data for each document in our index M2. We must infrastructurally support the size of our index and content stored in our database M3. We must offer a UI for searchers to select multiple facets M4. We must not introduce any security risks with these updates M5. The design must be accessible & 508 compliant M6. The design must be mobile-responsive M7. This feature must be usability tested M8. The facets feature must be Super-Admin settable	S1. We should capture analytics on facet usage S2. We should display the facet "matches" on an individual result S3. We should allow a user to easily remove facet selections, in the case they over-filter their query S4. Administrators should be able to turn the facets functionality on and off S5. We should add the count of documents that match category next to the facet option S6. Administrators should be able to rename facet labels to customize for their own business purpose S7. Administrators should be able to reorder the facet display for their results pages S8. We should allow a finite set of different data types (ex. string, list, date, number). List TBD. S9. We should gracefully handle missing data	C1. We could impose a limit on the number of facet values, or accept all. C2. We could allow customization on sort order of facet values (alphabetical or by result count) C3. We could persist the facet selections as query parameters, so users can directly link into queries with pre-selected facets C4. We could normalize / clean up facet values C5. We could offer a way to make facet values more plain-language	W1. We will not incorporate location-based or geolocation filtering W2. We will not attempt to provide facets for unstructured data W3. This will not be supported on Bing search experiences

	S10. We should limit the number of facet fields available for usage S11. This should be accessible to our API users		

MVP Requirements/features in: These are the distinct, prioritized features along with a short explanation as to why the features are important.

M1. We must have the ability to ingest structured data for each document in our index

In order to power the metadata driven search, we will need to capture this information in our index in a structured way. This should be a finite set of structured fields in our index, and ideally leverages metadata (Schema, OG, Twitter).

After an investigation (SRCH-2914) into the existing metadata in our index, we discovered the following:

- Aside from date fields, OG and Twitter may not be super helpful to us. They don't offer a lot of out of the box "categories" for things like Type, and for fields like keywords & tags there isn't a lot of adoption in our customer base
- There is some evidence of customers using DublinCore for tag-like content (dcSubject, dctermsKeywords)
- There is some evidence of customers using both DublinCore and Schema.org for AudienceType tagging
- Generally, the customers who are implementing Schema.org seem to know what they're doing; so the complexity may not be a significant barrier

This analysis and our customer requests over the years inform our proposed V1 fields: Faceted Search - Metadata Collection V1 Fields

M2. We must infrastructurally support the size of our index and content stored in our database

Given the expansion of our index, and the potential to onboard net new customers with this functionality, we need to ensure our ElasticSearch indices and other infrastructural components are able to withstand the size. At this point, we aren't concerned about the number of additional fields in ElasticSearch.

With this effort, we also may need to store our documents in our DB directly. This may have

additional infrastructure costs associated with storing this information. This will need a discovery effort.

M3. We must offer a UI for searchers to select multiple facets

We will build for selections to be “AND”ed between facet fields, and “OR”ed within a facet field.

M4. We must not introduce any security risks with these updates

All content that will be captured will be publicly accessible and seen in the HTML of these webpages, but we should ensure we’re not opening ourselves up to any security vulnerabilities in any way.

M5. The design must be accessible & 508 compliant

We should take lessons from existing compliant designs, as well as fully test through every step of the process the accessibility of the design. We will need to extend the automated accessibility testing tools implemented with the SERP Redesign to accommodate any changes made in the UI for faceted search.

M6. The design must be mobile-responsive

Although the majority of our users navigate the display on desktop (unlike most modern websites), we must support an intuitive filter design for smaller breakpoints (ie. tablet and mobile).

M7. This feature must be usability tested

M8. The facets feature must be Super-Admin settable

Super Admins should be able to turn the facets functionality on and off.

From Martha: That will greatly facilitate the testing workflow at every step of development. (It could also help to have a "facets_enabled" affiliate seed). Note also that we'll need higher-level logic than a boolean field for "facets_enabled", as it will have to be off for Bing or I14y affiliates, no matter what the boolean db value is.

S1. We should capture analytics on facet usage

Just like any component of search, it is only useful if people using it are better able to find the results they need. To quantify this, we will need to capture several metrics related to facets, on both the query and aggregate level:

- CTR of searches using facets vs. not
- Most used facet fields (Counts)
- Most used facet values (Counts)

S2. We should display the facet “matches” on an individual result

Users need to know *why* a result is displaying, so showing what filters it matched to is important.

S3. We should allow a user to easily remove facet selections, in the case they over-filter their query

Users may inadvertently end up over-filtering their query. We should give users a way to remove filters and search for their search term again.

S4. Administrators should be able to turn the facets functionality on and off

Just like any of our other search features, this should be easy to toggle off and on.

S5. We should add the count of documents that match category next to the facet option

This informs users on the expected quantity of results after they apply filters.

S6. Administrators should be able to rename facet labels to customize for their own business purpose

The facet labels should reflect how a site’s searchers navigate the content, so should be easily modified to resonate with their audience.

S7. Administrators should be able to reorder the facet display for their results pages

Site administrators may have a preference for the order the facet fields display in, given they’re available for the result set returned.

S8. We should allow a finite set of different data types (ex. string, list, date, number). List TBD.

This will impact the UI considerations for how we handle data types. For V1, we propose we support string (and list) and date filtering.

S9. We should gracefully handle missing data

We will likely encounter sites that have metadata inconsistently applied across their document base. We want to ensure this content still appears if relevant, but also make it clear where categorization is missing. This could be a “Uncategorized” facet value that appears under each facet field as the last result.

S10. We should limit the number of facet fields available for usage

We want to ensure that the facets that are added to the experience are meaningful to users, and that the UI can support the number of fields our administrators choose to feature.

S11. This should be accessible to our API users

To encourage feature parity, and because some of the important requestors of the facet functionality are API users (VA, NICHD/NIH), we should expose this functionality in the API.

C1. We could impose a limit on the number of facet values, or accept all.

Depending on what we end up indexing per document, we may want to impose a limit on how many facet fields can be added, or a limit on how many may display in the UI. This will also have UX implications.

C2. We could allow customization on sort order of facet values (alphabetical or by result count)

This would get tricky, so likely should be a V2 if it's requested by customers. Default to start will be by result count.

C3. We could persist the facet selections as query parameters, so users can directly link into queries with pre-selected facets

This is commonly supported in other faceted search interfaces, but we should scope technical feasibility of supporting this.

C4. We could normalize / clean up facet values

There may be different variations within facet values passed to us (based on casing, misspellings, and other automatically-detectable differences) that we could consolidate into one facet field. We likely will need to do this to Date fields in order to present accurate date ranges when applied, but could defer this for text-based fields.

C5. We could offer a way to make facet values more plain-language

Best practices suggest that facets are only useful if they reflect the language users use to search their content. However, this may not be how content is currently tagged within the webpages in our index. We may want to offer a way to update the facet values within our system so that they are more user friendly.

MVP Features out: What have you explicitly decided are post MVP items and why?

W1. We will not incorporate location-based or geolocation filtering

This data type will be complicated to handle in a V1, and demand may not justify investing

the development resources at this time.

W2. We will not attempt to provide facets for unstructured data

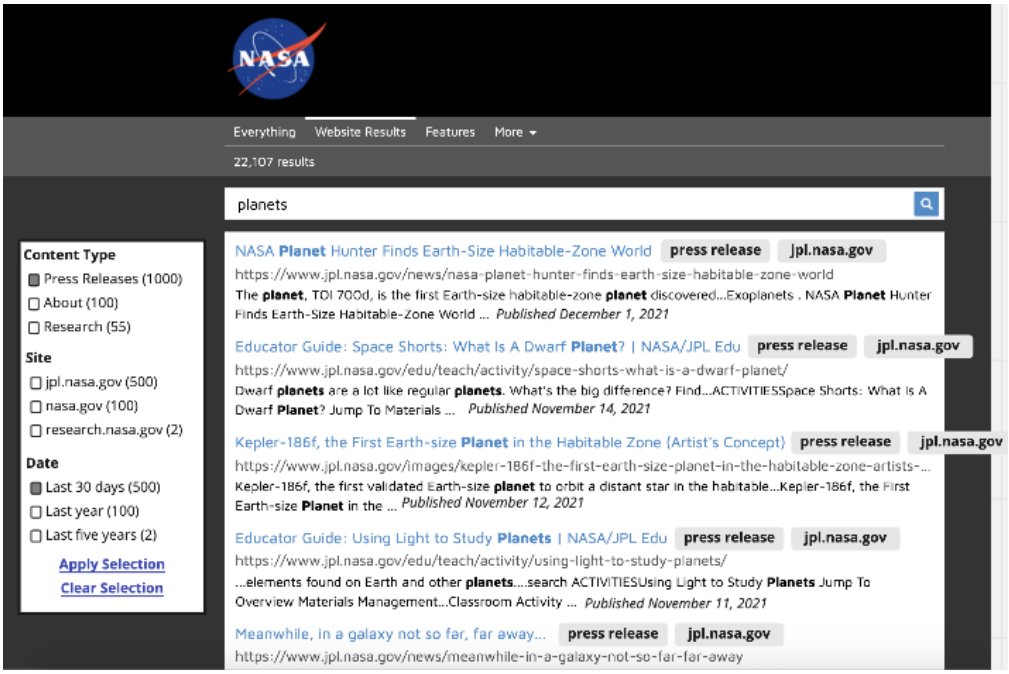
There are ways we can extract entities & themes from unstructured data, but this is not in scope for our initial release. If there is significant demand in the future, we may reconsider.

W3. This will not be supported on Bing search experiences

Much like other features offered by our service, we will require customers to be on our index to implement facets.

Designs: Include any needed early sketches, and link to the actual designs once they're available.

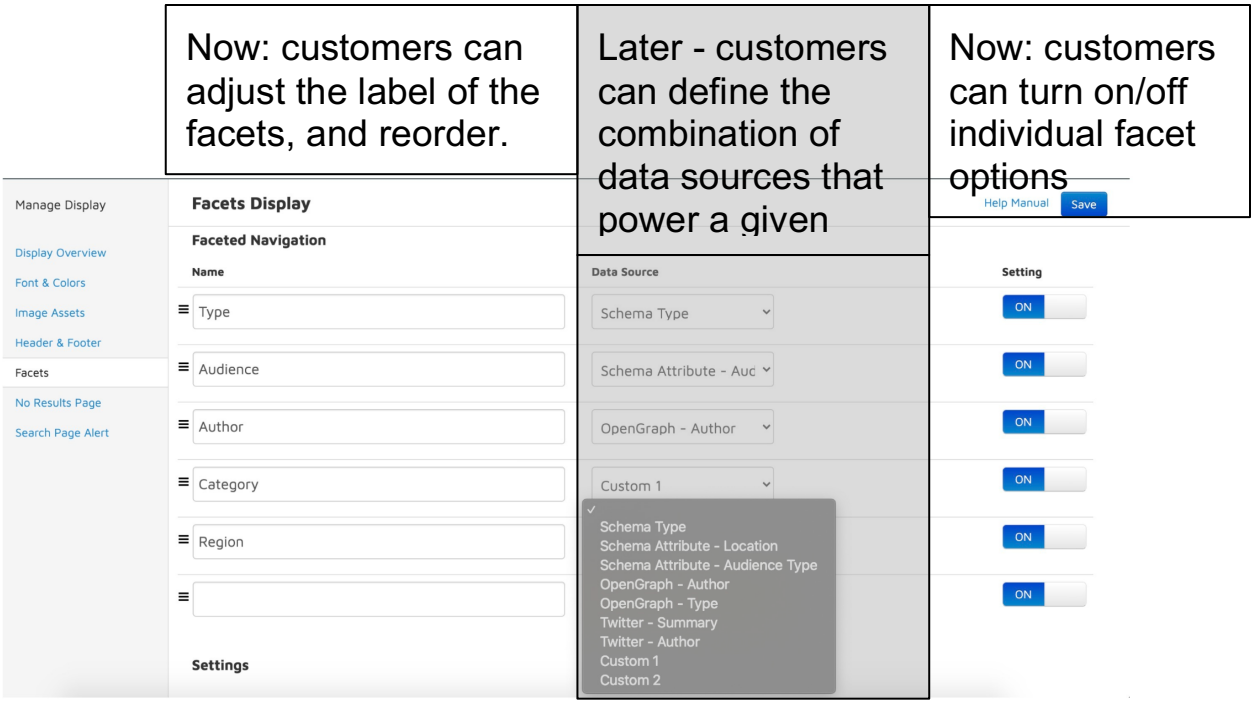
Search UI



Rough initial sketch to illustrate components

Admin UI

FOR DISCUSSION PURPOSES - NOT PRESCRIPTIVE APPROACH



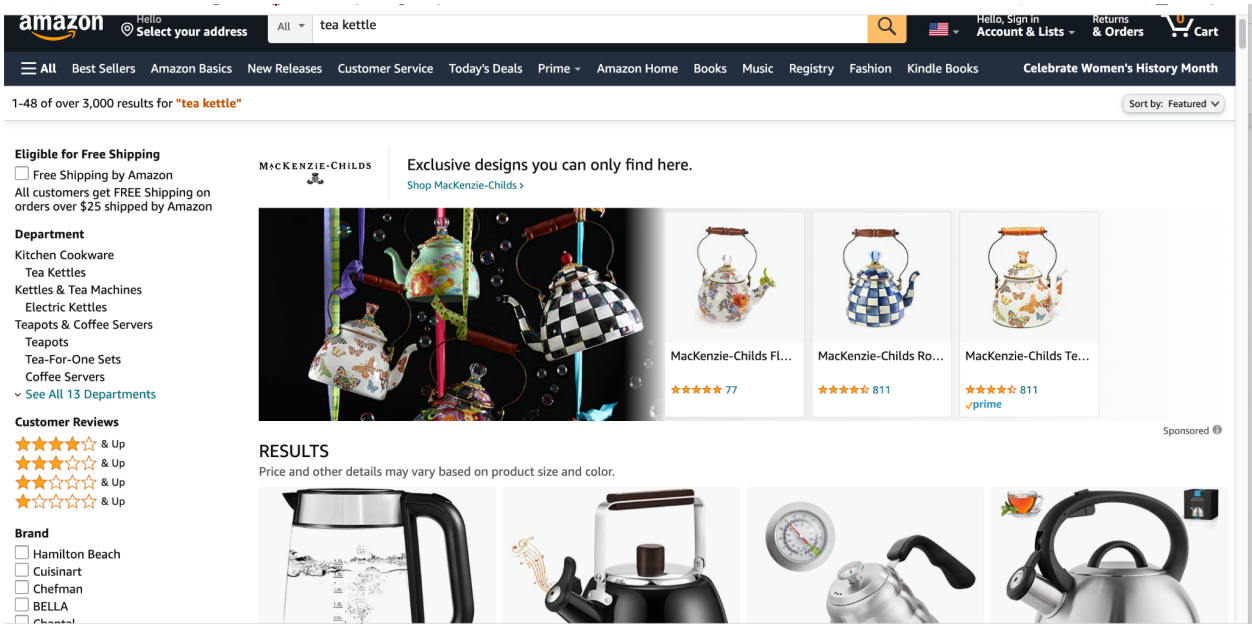
Rough sketch of Facet Configure screen

Open issues: What factors do you still need to figure out?

Current risks and open issues can be found here. This is a living list that is constantly evolving.

- What dependencies does this have?
    - Results page redesign?
    - ElasticSearch upgrade?
    - Advanced Search functionality?
    - Other code zombies in our way?
- How should we communicate what *we* will use from existing markup (Schema.org, OpenGraph, etc), and what would be best recommendations for Google/organic search?
- How should we show the metadata that matches the applied filter on a specific result? Ex. if someone selects that they want to see Press Releases and Biographies, how should we display the category a result belongs to on the snippet itself?
    - The "Tag" USWDS component may be helpful here
- The facets will only impact the organic search / Garfield module. Should we only offer the faceted search UI on a vertical page (and have a link out from the main search page to take them to the vertical)?

    - This follows the paradigm of other federated search engines - ex Google really only offers facets on their vertical searches (images, shopping, flights) but not on the main search.
- How will we handle presenting non-web result content alongside filtered results?
    - Different “Web” vertical that supports filtering, instead of on the “Everything” tab?
    - Assuming that Twitter, Youtube, RSS Feeds, Images, Best Bets, etc are out of scope for this filtering
    - “Everything” vertical could have everything, like in this screenshot:



    - As soon as a filtering option is selected, could clear out all the non-web modules. Worth noting that Amazon keeps the ads around though at least the first filter.
- What kinds of documents will we support in faceted search? How would we get metadata for docs like PDFs, .txt, Excel, etc?

Q&A: What are common questions about the product, and answers to those questions? This is a good place to note key decisions.

Q: How will we incorporate our customers’ feedback in the development process?
A: We will be hosting a few focus group in January to better understand how customers are structuring this data on their end, and how to facilitate the handoff of that information to our indices. We also will be identifying some early partners who can help us test the indexing mechanism as well as the initial UI.

Q: Collections-Facets interaction:

- **Can we easily use the Collections feature as values in a facet?**

- **Are facets going to appear on the /docs vertical, or just on the /search vertical?**
- **Customers want to be able to scope collections in the admin center using metadata values, in addition to paths. Is this in scope of this work?**

A: We are not considering offering collections as facet values at this time, since it will require a significant refactoring of our existing index. Expanding the definition of collections outside of folder paths to include metadata is also out of scope, though certainly better enabled by the metadata we will be collecting.

Q: How does this feature interact with the existing advanced search functionality? EX: https://search.usa.gov/search/advanced?affiliate=usagov
A: The advanced search form only is available for Bing users. It will not be updated as part of this effort, even though it is slowly being phased out. Because faceted search will only be available for those on the Search.gov index, there should be no interaction between the two components. Further thought into how we handle the Bing Advanced Search will be part of the Advanced Search effort (PRD).

Appendix

Previous Product Opportunity Document

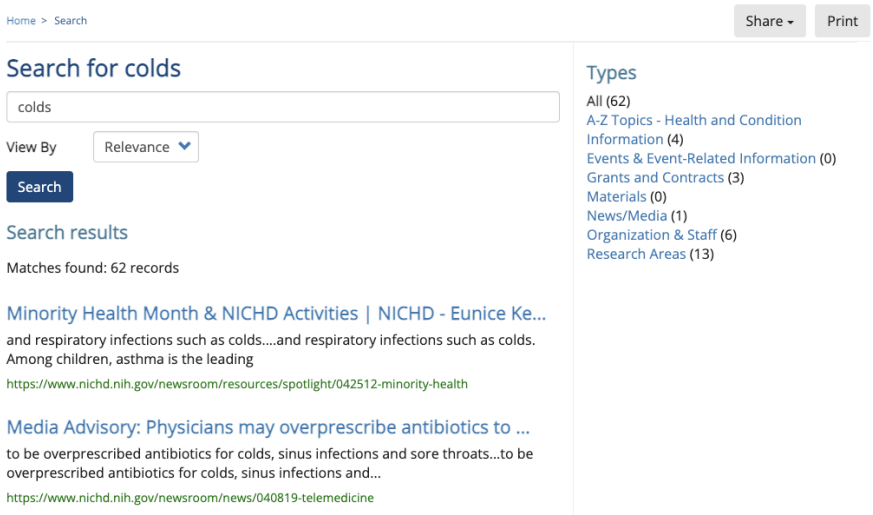
PIF Recommendations for Faceted Search Implementation

Research Components

- Faceted Search Research Plan
    - Literature Review (PIF Research)
    - Customer Surveys (2020 & 2022 Surveys, Quarterly Check-in Questions)
    - Focus Groups
    - Non-Customer Research
    - Google Search Console Analysis
    - Customer Requested Facets Analysis
- Existing Metadata Analysis (Summary, Raw Data)

Other Background

Some customers have developed creative solutions to work around this; for example, nichd.nih.gov has created separate search sites for each "category", and makes several simultaneous API calls to allow users to filter the set of returned results more easily.



Example from nichd.nih.gov

However, this has negative implications on their ability to view consolidated analytics, and increases the load on our system. Rather than requiring customers to find unique solutions or pursue a different search solution entirely, we want to offer this functionality as a feature of using Search.gov.