OBP

Phase 2 Report

Improvements to the Pilot Application

*Updated to include follow-on work in Dec 2018 & Jan 2019.*

January 31, 2019

\_\_\_\_\_\_\_



[**Background**](#_s25pcmxlub96) **3**

[**Module 1 Updates**](#_w1t7trdviy9y) **5**

[Related Terms](#_3n0j8hahvhmi) 5

[Metadata](#_71fyijo166zq) 5

[DSpace Sync Frequency](#_rggtzvtc0u9a) 6

[Tag Relationships](#_6198sz817p9m) 7

[OceanKnowledge Tagger](#_ggo90iqkpaj3) 7

[**Module 2 Updates**](#_cyc5tvq9hc93) **7**

[Related Terms in Lookahead](#_2tco4c3elc2l) 7

[Deeplinking](#_fye106pi06wt) 7

[Terminologies (Ontologies and Vocabularies)](#_pl7xpmnt7lac) 8

[**Module 3 Updates**](#_gz4iwp8h3v2l) **8**

[Minor UI Improvements](#_awrqd7q6g226) 8

[Landing Page](#_1f5ghhmd2x6q) 9

[Search Bar](#_ynu1zssicm7w) 9

[Search Tips](#_vrcgc43i1v7a) 11

[Search Results Presentation](#_drtz8mk1yqrx) 12

[Tag User Interface](#_q0hhxyaofx4) 12

[Query Presentation and Manipulation](#_9ar11m726bzq) 15

[PDF Viewer With Term Navigation](#_5rv77ychgd83) 16

[OceanKnowledge Tagger UI](#_p0n2o9pfyhs9) 17

[**Google Analytics Integration**](#_9vmgbe7d19os) **19**

[Basic Setup](#_hnzce3vrixv1) 19

[Viewing and Analyzing Data](#_qtj1jzimpwwt) 19

[Finding the OBP Google Analytics Dashboard](#_yd096mdc7jl1) 19

[How do people search on the website?](#_of0z5u8ku0aw) 20

[What do people interact with on the website?](#_axskwp3w7opf) 21

[What kind of demographics and technology visits the website?](#_m5hdetiaqa24) 21

[More In-Depth](#_ta1fcjb5xa2z) 22

[**OceanKnowledge Tagger API**](#_rp6vzrtq973h) **23**

[**Backup and Recovery Plan**](#_fmrworr1n6dw) **24**

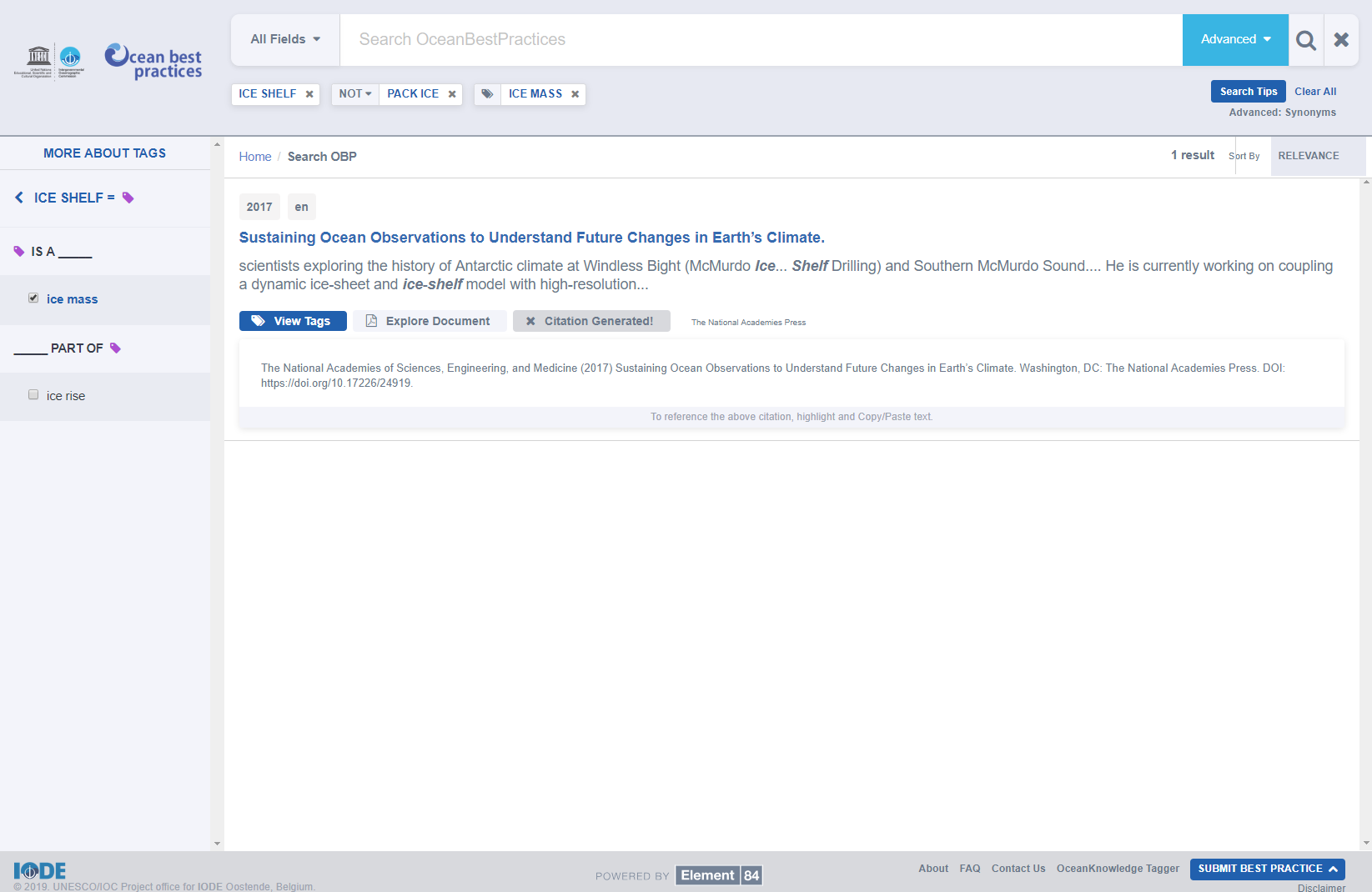
[Source Code](#_ec5yhyyzex57) 24

[Project Infrastructure and Data](#_j6qo9awk1oq3) 25

[**Appendix 1: System Architecture**](#_tgu20pjzip5s) **26**

[**Appendix 2: Google Analytics Deep Dive**](#_2gel19z6b083) **27**

# 



# Background

The OceanBestPractices Repository (OBP-R) is a DSpace-based repository providing simple search capabilities. The goal of this effort is to move toward a more powerful, semantic search tool for users to more effectively find documents within the repository.

In Phase 1, Element 84 was tasked to create an application that ingests new documents found in the OBP-R, maps these documents against specified ontologies, and provides an interface allowing users to search ingested documents in a semantic fashion. Details on Phase 1 are found in a separately issued report.

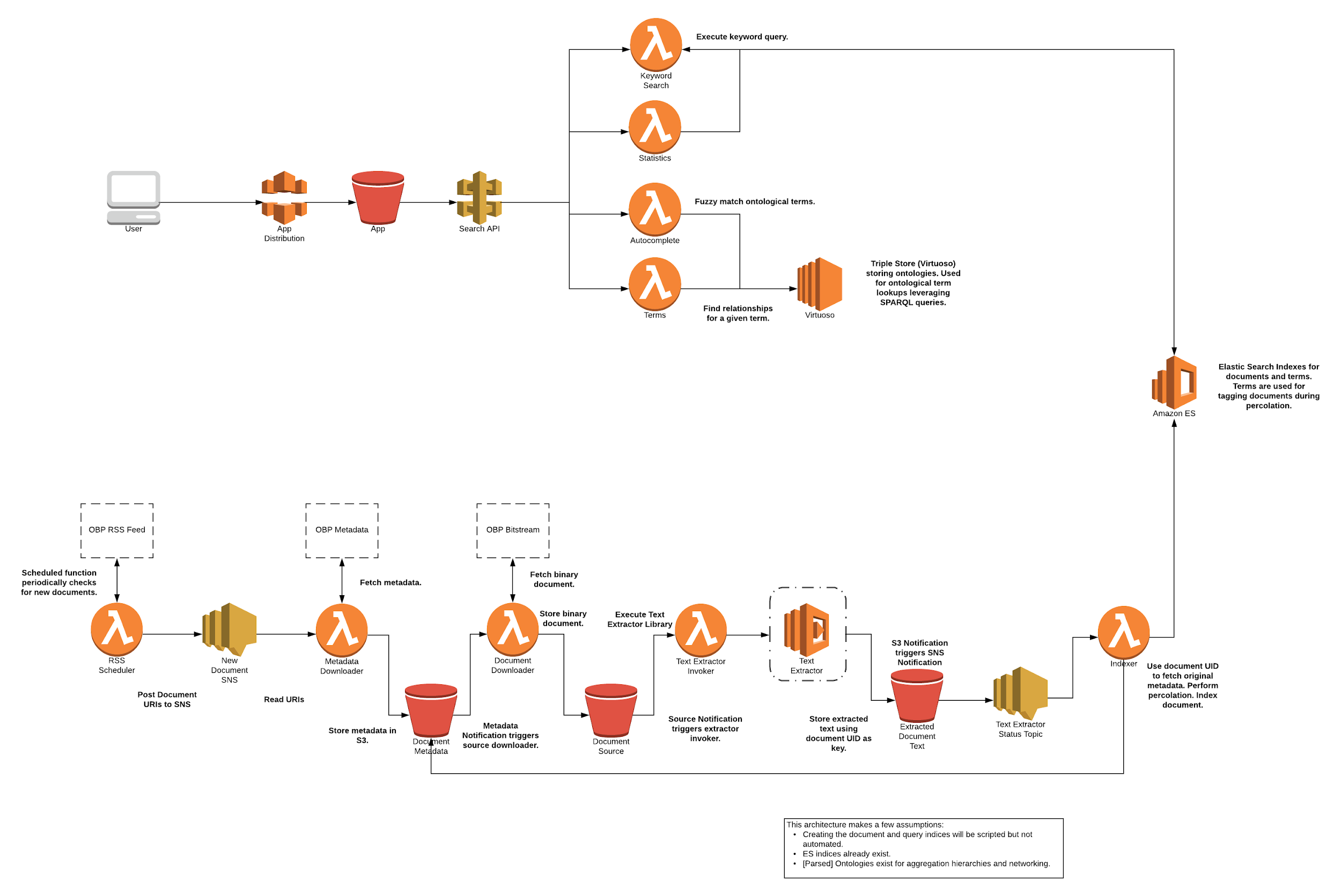
Phase 2 continues this effort with a focus on improvements that give users more tools to find best practices.

The three parts to this application are:

**Module 1**: Semantic tagging of OBP-R documents

**Module 2**: Semantically enhanced search and browse capacities

**Module 3**: User Interfaces



**FIGURE 1:** Overall system architecture (see Appendix for larger version).

# 

# Module 1 Updates

## Related Terms

Keyword search and autocomplete respect a user defined option that can include synonyms (e.g. exact synonyms; alternative labels; related annotation properties) when querying for documents or executing autocomplete of a search term. This feature is exposed to the user via an Advanced Search option. When enabled, the search API queries ingested ontologies and vocabularies for related terms. These related terms are included in the Elasticsearch query or returned directly to the user in the case of autocomplete.

## Metadata

The dc.identifier.citation field is now indexed as part of the document metadata during ingest. We expose this metadata field to the user via a “Generate Citation” button located in the search results. The user can directly copy this citation for use in other works.

Several metadata fields have been added to support user-specified metadata field searches. Metadata for all documents, whether or not that document has a PDF bitstream, is indexed. The full list of indexed fields is below:

|  |  |
| --- | --- |
| **FIELD** | **USE** |
| dc.date.issued | Search results presentation |
| dc.title | Search results presentation  Metadata search: Title |
| dc.publisher | Search results presentation  Metadata search: Publisher |
| dc.contributor.author | Search results presentation  Metadata search: Author |
| dc.language.iso | Search results presentation |
| **FIELD** | **USE** |
| thumbnail | *No longer in use* |
| uuid | Site navigation |
| handle | Site navigation |
| dc.title.alternative | Search results presentation |
| dc.contributor.corpauthor | Search results presentation |
| dc.contributor.editor | Search results presentation |
| dc.bibliographicCitation.title | Search results presentation |
| dc.description.eov | Metadata search: EOV |
| dc.description.sdg | Metadata search: SDG |
| dc.identifier.doi | Metadata search: DOI |
| dc.resource.uri | Site navigation |
| dc.description.refereed | Advanced search: Refereed filter |
| dc.identifier.citation | Search results: Generate Citation |
| dc.description.abstract | Search result boosting |

## DSpace Sync Frequency

The system now requests the DSpace RSS feed every 5 minutes, comparing the publication date of the feed to the previously known update time. The rate at which the RSS feed is checked is configurable. *Note: the DSpace RSS feed (beyond the scope of this project) ought to be updated to match the 5 minute interval to take reduce delay of new best practices being indexed by Module 1.*

## Tag Relationships

The API for requesting a tag neighborhood has been updated to provide directionality (e.g. parent, child, sibling context) in the response. This allows for any client to determine how tags are related and present this relationship to the user.

## OceanKnowledge Tagger

The OceanKnoweldge Tagger API (OKT API) provides a service to users where they can send raw text and receive a list of terminology labels that, according to the rules configured as part of the OceanBestPractices service, match the received text. A separate section later in this report goes into further details on this API.

# Module 2 Updates

## Related Terms in Lookahead

If the user has enabled the Synonyms Advanced Search option, an extra parameter is sent to the API as part of the autocomplete (i.e. lookahead) request. This ensures that synonyms in addition to fuzzy matches are included in the autocomplete drop down.

## Deeplinking

Users can access specific search results using a link which provides the context necessary for the system to execute a search. Users can share this link with others for quick discovery of a specific search.

Example: <https://www.oceanbestpractices.org/search?q=Law%20of%20the%20sea,%2Bfish%20processing&fields=all>

## Terminologies (Ontologies and Vocabularies)

The following ontologies and vocabularies have been added and made searchable as part of this update:

* SDGIO (OWL)
* L05 - SeaDataNet device categories (SKOS)
* L06 - SeaVoX Platform Categories (SKOS)
* L22 - SeaVoX Device Catalogue (SKOS)

This brings the total terminologies to 6 (ENVO and CHEBI being added previously).

# Module 3 Updates

## Minor UI Improvements

A number of changes and improvements were added following a first round of user feedback:

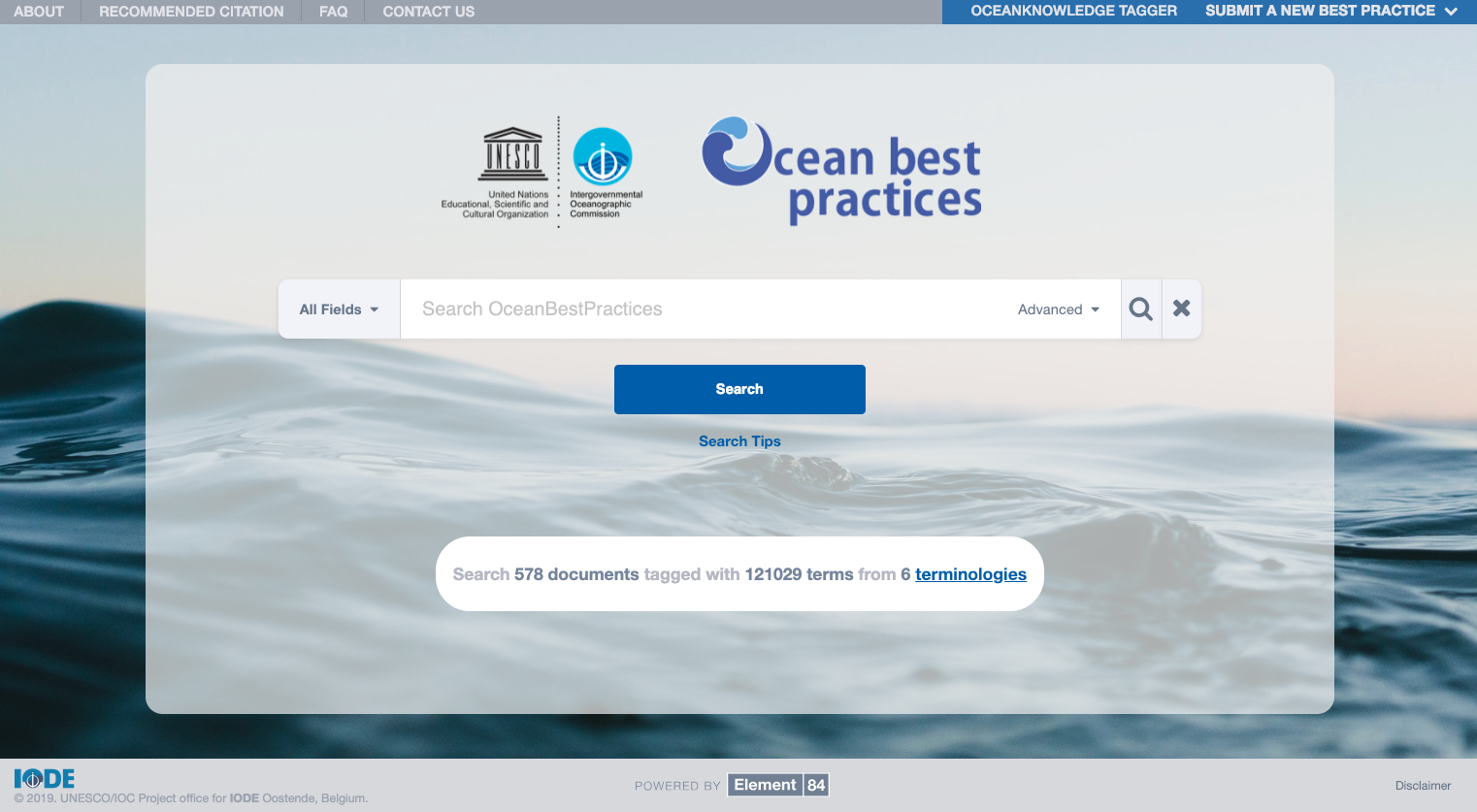
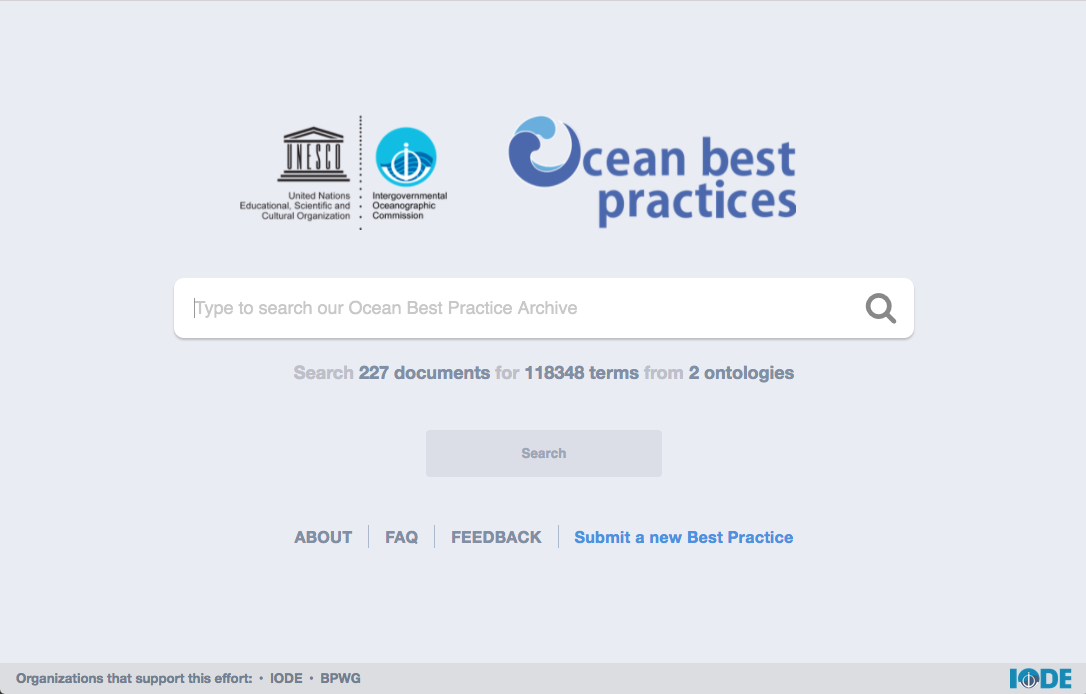
* Recommended Citation was added as a link on the landing page header and describes to users how to cite this tool.
* A list of templates to be used for best practice submission were added under the “Submit A New Best Practice” menu in the landing page header, and search results footer.
* Links to the OceanKnowedge Tagger were added to the landing page header and search results footer (more details on the OceanKnowledge Tagger follow).
* Search and Clear Search buttons persist in the search bars
* Clear All now resets the search to All Fields as well
* Order of fields presented in field dropdown was adjusted per feedback
* The “More About Tags” link to Search Tips is now persisted in the left Tags panel
* “Search Tips” on the Search Results page has been outlined for more visibility

## Landing Page

The landing page has been updated to feature a background image, include the ‘Search Tips’ modal, and move the links underneath the search bar to the header. The landing page will scale to all device sizes (mobile, tablet, desktop). As a whole, the landing page remains the same where it still allows users to begin their search by entering a query in the prominently displayed search box.

The links on the footer have been removed, the IODE logo has been moved, the Element 84 logo has been added, and the copyright has been altered to update with the current year.

“Ontologies” has been replaced with “Terminologies” to more accurately represent the mix of ontologies and vocabularies used by the system. Clicking on “Terminologies” reveals a hard-coded list of terminologies used. This list will need to be updated as new terminologies are added or eventually changed to automatically update.

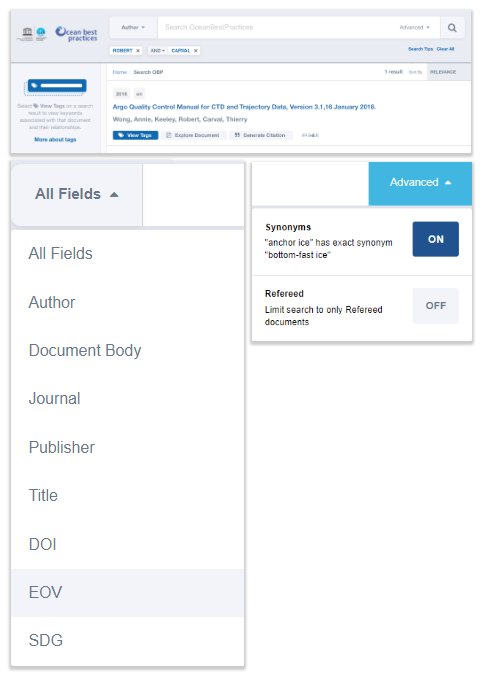


**FIGURE 2:** Old Landing Page **FIGURE 3:** New Landing Page

## Search Bar

The search bar has been updated in in a number of ways. Users are now able to execute search queries over metadata fields (fig 4). These metadata fields are from the OBP-R archive, as well as the tagged corpus. An example of this is if a user were to select the metadata field ‘Author’, and search for ‘Robert AND Carval’, the search results would include documents that have Robert and Carval as authors (fig 4).

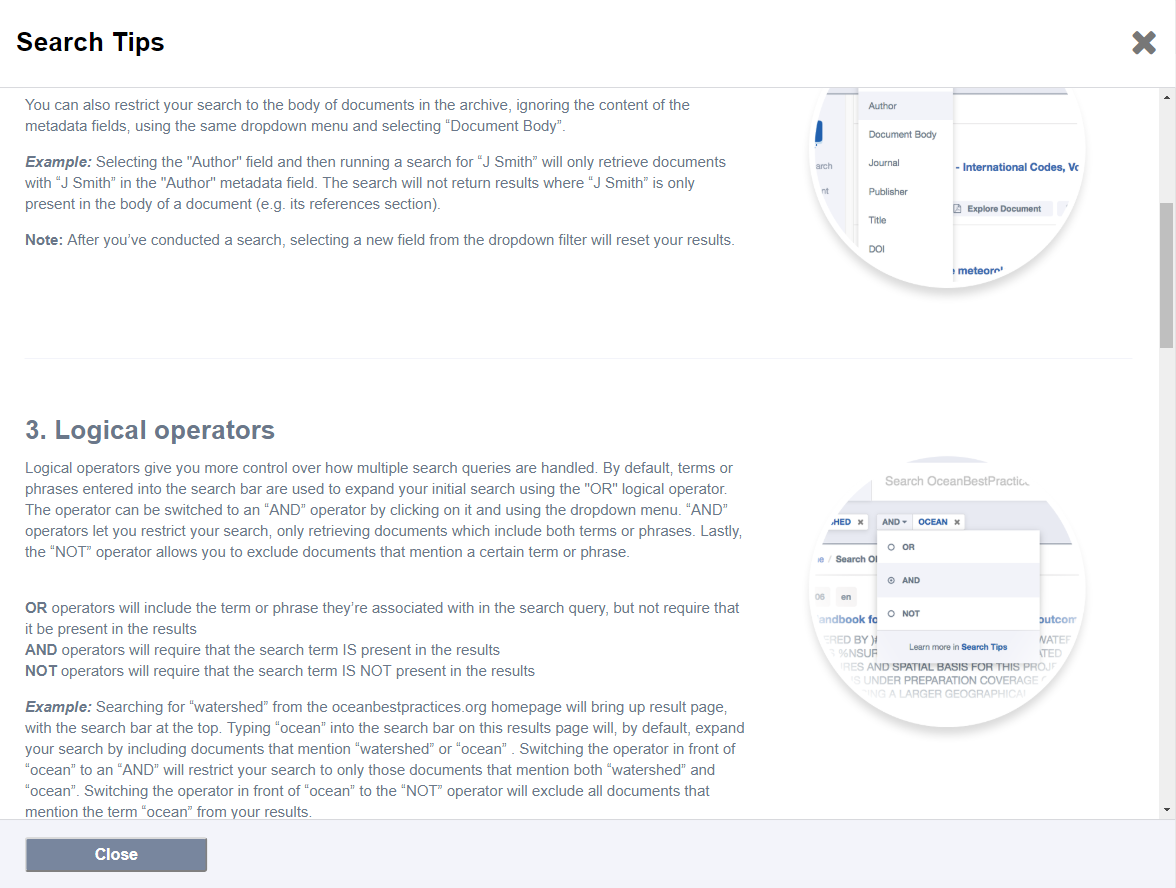
Toggling the ‘Synonyms’ field will include exact synonyms in the query. For instance, ‘anchor ice’ has the exact synonym of ‘bottom-fast ice’. The user can also toggle the ‘Refereed’ field, which will limit the search to only refereed documents. As a whole, toggling an advanced field will help to narrow down the search.



**FIGURE 4:** Updated Search Bar: Field Selection and Advanced Options. Note: field ordering has changed due to user feedback.

## Search Tips

A ‘Search Tips’ modal has been added. The search tips are helpful explanations about building and manipulating queries. At the bottom of the ‘Search Tips’ modal, there will be an embedded YouTube player. When the YouTube ID is added to the ‘youtubeVideoID’ constant in the ‘ src/js/components/SearchTipsModal.js’ file, the YouTube player will render, if there is no YouTube ID present, the player (and the search tip the video tutorial is in) will not render. When the user clicks on *Search Tips* a modal dialogue opens and can be closed by either clicking on the ‘X’, the ‘Close’, or outside of the modal. The modal is scrollable. There are several links to the search tips. The links can be found on the landing page, underneath the advanced search bar, and in the side panel.

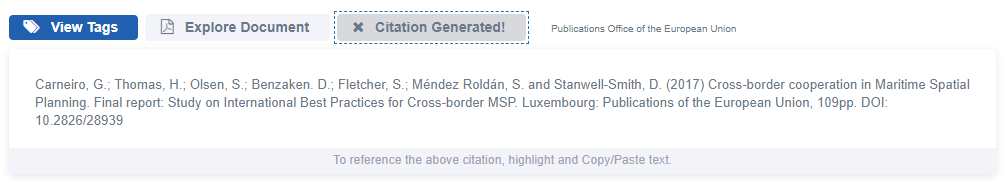


**FIGURE 5:** Search Tips

## Search Results Presentation

Search results have been updated in several aspects (fig 6). The thumbnails have been removed. If a document has an author (or authors), it will list them underneath the title of the document. The query shows up in the highlight with preceding and anteceding ellipsis (instead of commas). The publishing location has been removed, and instead only the publisher is shown. Several new buttons have been added as well. The ‘View Tags’ button remains the same. The ‘Explore Document’ button will populate a PDF viewer with the query. The ‘Generate Citation’ button only appears on documents where a citation is available. Upon clicking the ‘Generate Citation’ button, users will be presented with a text box where they can highlight and Copy/Paste the citation (fig 6).

**FIGURE 6:** Updated Search Results UI

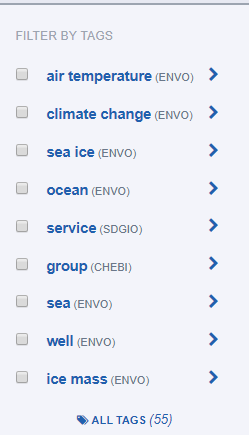
**FIGURE 7:** Citation Generated

## Tag User Interface

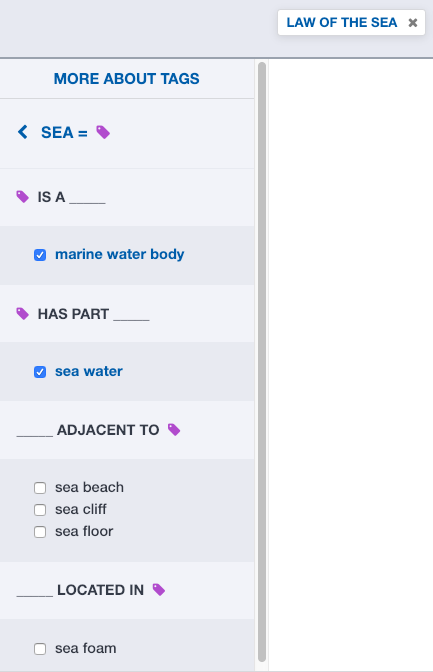
Selecting the ‘View Tags’ button in the search results will populate the document tags panel. The top level tags are now selectable and adding them to the query (selecting the checkbox) will further filter search results. Multiple top level tags can be selected. Like before, selecting a top level tag (not selecting the checkbox, but the tag itself) will show the various relationships associated with that tag.

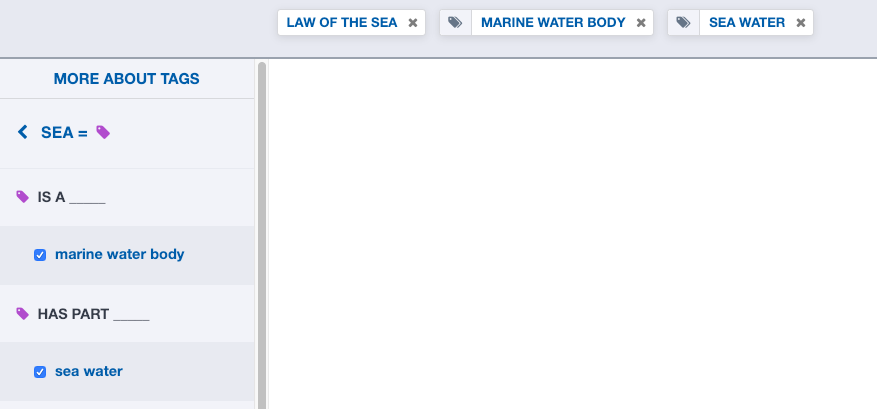
A tag icon and an underlined space have been added to represent the directionality of the tags while looking at the various relationships associated with the tag. For instance, (in fig 9), it reads ‘Sea’ is a ‘marine water body’ because the tag icon (‘Sea’) is before the ‘is a’ and is followed by the underlined space (‘marine water body’).

When selecting a tag, the tag will populate underneath the search bar (fig 10). A user can select the tag to remove it, and subsequently remove the tag from the query. The main search (from the search bar) will also populate underneath the searchbar, and the same functionality applies. Tags are represented with a tags icon, whereas the main search will not have any icons.



**FIGURE 8:** Filter by Tags

  
**FIGURE 9:** Tag Relationships



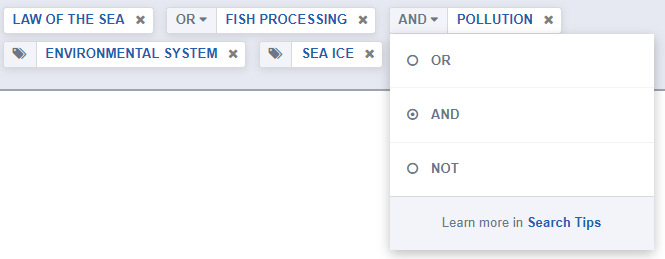
**FIGURE 10:** Tags populate under Search Bar

## Query Presentation and Manipulation

Users can now use logical operators in their query. These logical operators are: OR, AND, and NOT. The ‘Search Tips’ modal includes help on how to use the logical operators. Additional searches are added to the existing query, instead of replacing the existing search. For instance, if a user were to search for ‘Law of the Sea’, a tag will populate underneath the search bar with ‘Law of the Sea’ and the search bar would be empty. If the user were to search for ‘Fish Processing’, an additional tag would populate next to the already existing ‘Law of the Sea’ and the search bar would again, be empty. All searches following the initial search are associated with a logical operator (fig 11). The search term, as well as the tags, can be clicked on to be removed. Additionally, there is a clear all (fig 4) button that, when selected, will clear all the tags and reset the query.

As stated above, search terms do not have an icon associated with it, while tags do (fig 11). For example, ‘Law of the Sea’ does not have an icon, where ‘Environmental System’ does.

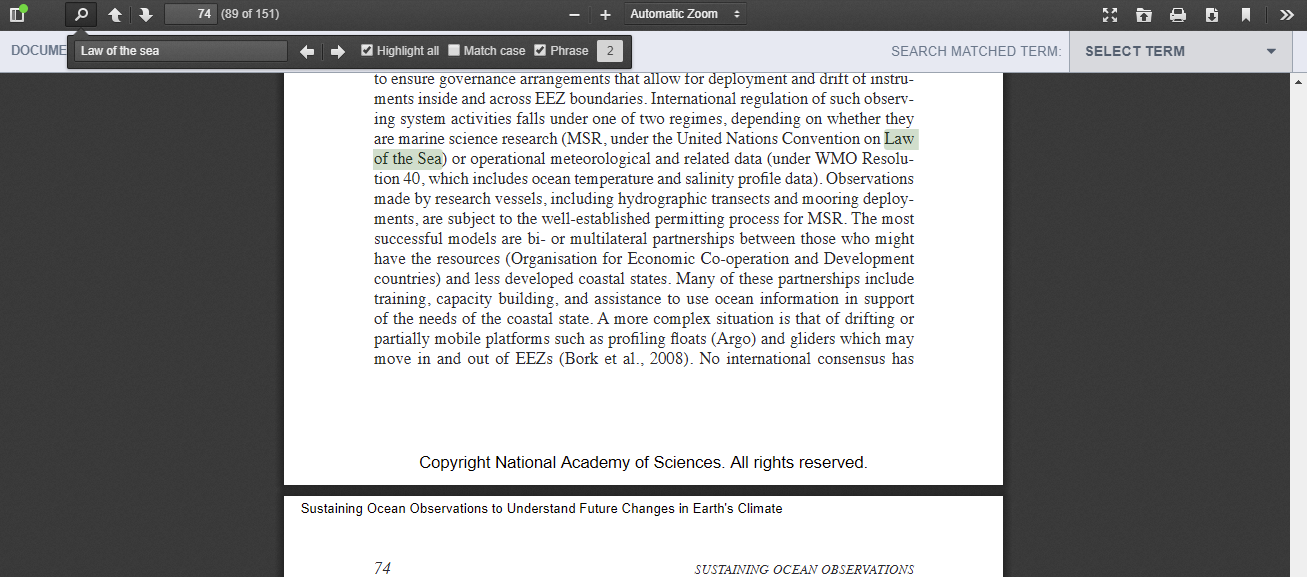
The logical operator menu allows users to select a logical operator. The default logical operator is OR. Selecting the logical operator next to the search tag will open up the logical operator menu, where a user can select another logical operator. Selecting an operator or clicking outside of the menu will close the menu.



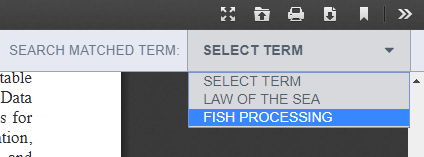
**FIGURE 11:** Logical Operators used with Search Terms

## PDF Viewer With Term Navigation

Selecting the ‘Explore Document’ button on the search results will open and populate a PDF viewer with the location of the query highlighted in the PDF. From there, the user can ‘jump’ to where the query shows up next in the PDF by selecting the arrows in the header (fig 12). A user can navigate forwards and backwards between query locations in the PDF. If the query was for ‘Law of the Sea’ OR/AND ‘Fish Processing’, then the user would be able to change the search term in the PDF viewer by selecting the dropdown ‘Select Term’ (fig 13). The user can also download the PDF from the viewer, as well as print the PDF. There is a known issue where a multi-word search term is showing up in the search results, but is  
not showing up in the PDF because of extra spaces in between words in the PDF due to  
PDF justification.

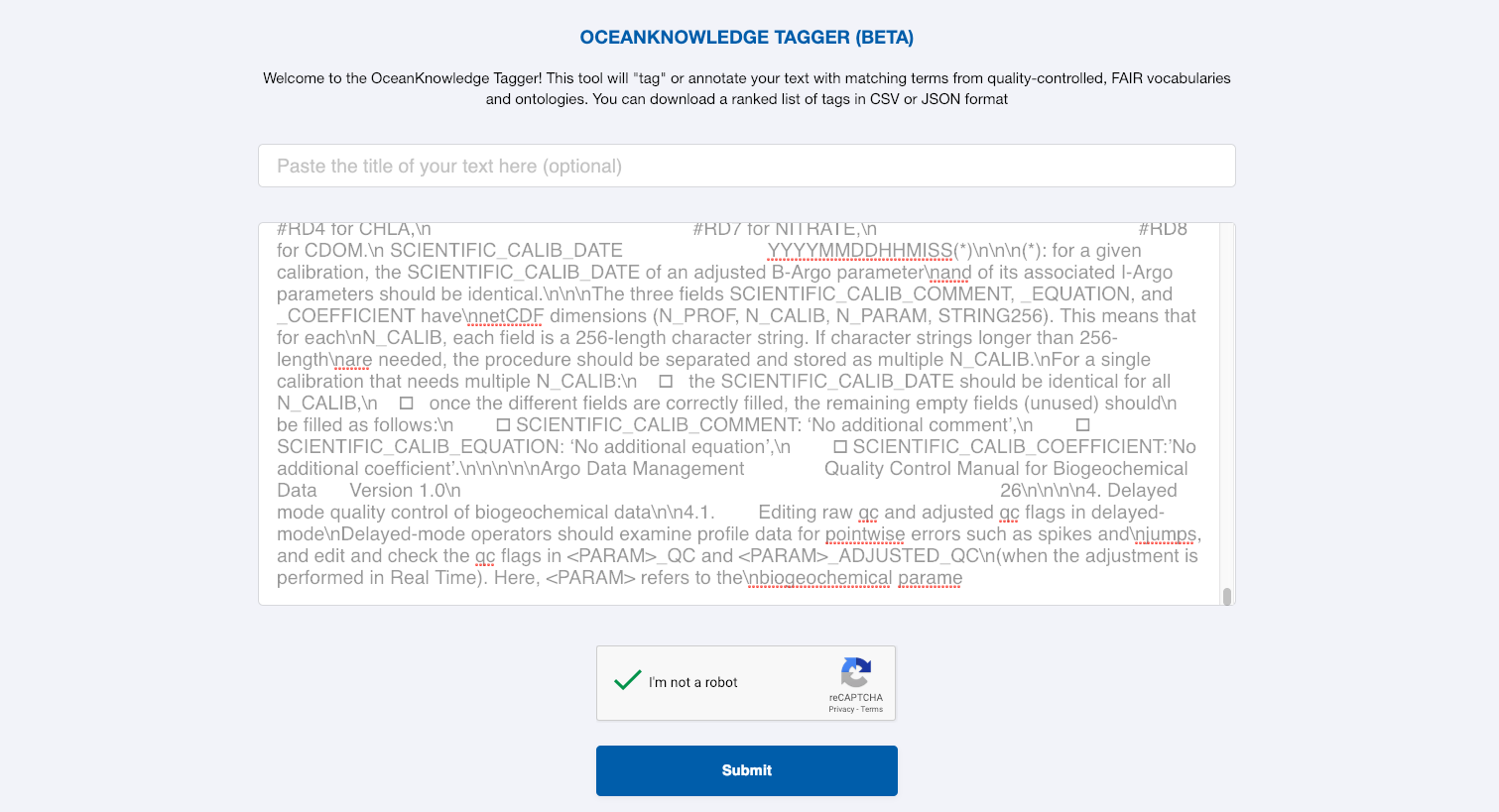


**FIGURE 12:** PDF Viewer

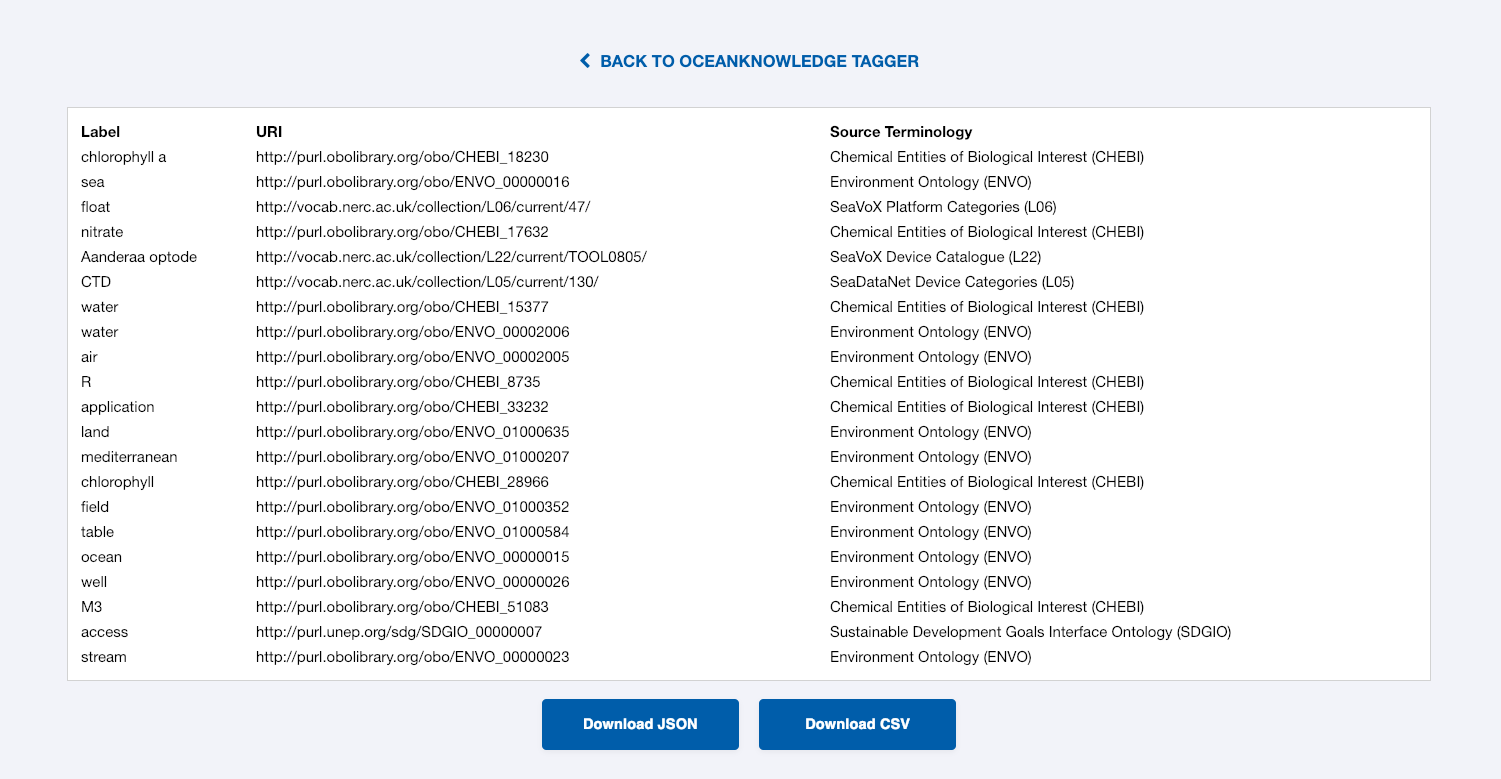
  
**FIGURE 13:** Term Navigation

## OceanKnowledge Tagger UI

The OceanKnowledge Tagger is where a user can enter raw text (fig 14) and receive a list of terminology labels that, according to the rules configured as part of the OceanBestPractices service, match the received text (fig 15). The maximum amount of characters that can be entered is 50,000 (fifty thousand), and can be altered by changing the ‘contentMaxLength’ constant in the ‘src/js/layouts/Tagger.js’ file. The list of terminology labels can be downloaded as a .json or .csv. Google’s reCAPTCHA is implemented on this page to reduce abuse by automated systems.



**Fig 14:** OceanKnowledge Tagger UI Input



**Fig 15:** OceanKnowledge Tagger UI Results

# OceanKnowledge Tagger API

The OceanKnowledge Tagger API is an endpoint that is made available to clients as part of the OceanBestPractices API. The goal of the OceanKnoweldge Tagger API (OKT API) is to provide a service to users where they can send raw text and receive a list of terminology labels that, according to the rules configured as part of the OceanBestPractices service, match the received text.

The OKT endpoint exposes the same routines that are used by the OceanBestPractices indexing service and provide the user with a preview of which terms would match their document had they published it to the OceanBestPractices repository.

## Accessing the OceanKnowledge Tagger API

You can call the OKT API with the following command:

`curl -XPOST <https://oceanbestpractices.org/documents/preview>`

The request expects a JSON body with the string fields `title` (optional) and `contents`:

```

{

“title”: “An optional but useful title Volume 3”,

“contents”: “This is the document body.”

}

```

The OKT API will return a result which includes, by default, a JSON list of matched terms:

```

[

{

"label": "chlorophyll a",

"uri": "http://purl.obolibrary.org/obo/CHEBI\_18230",

"source\_terminology": "Chemical Entities of Biological Interest (CHEBI)"

},

{

"label": "air",

"uri": "http://purl.obolibrary.org/obo/ENVO\_00002005",

"source\_terminology": "Environment Ontology (ENVO)"

},

{

"label": "float",

"uri": "http://vocab.nerc.ac.uk/collection/L06/current/47/",

"source\_terminology": "SeaVoX Platform Categories (L06)"

},

....

]

```

The endpoint supports the following URL query parameters:

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Allowed Values** | **Default Value** | **Description** |
| form | json|csv | json | Determines the format in which the results will be returned. |

# 

# Backup and Recovery Plan

## Source Code

Element 84 uses common best practices when developing software. All project source code, examples, and resources are version controlled with *git* (<https://git-scm.com/>) and pushed to a remote repository managed by Element 84. The remote repository is backed up on at least a 30 day rolling schedule and can be restored within an acceptable time frame.

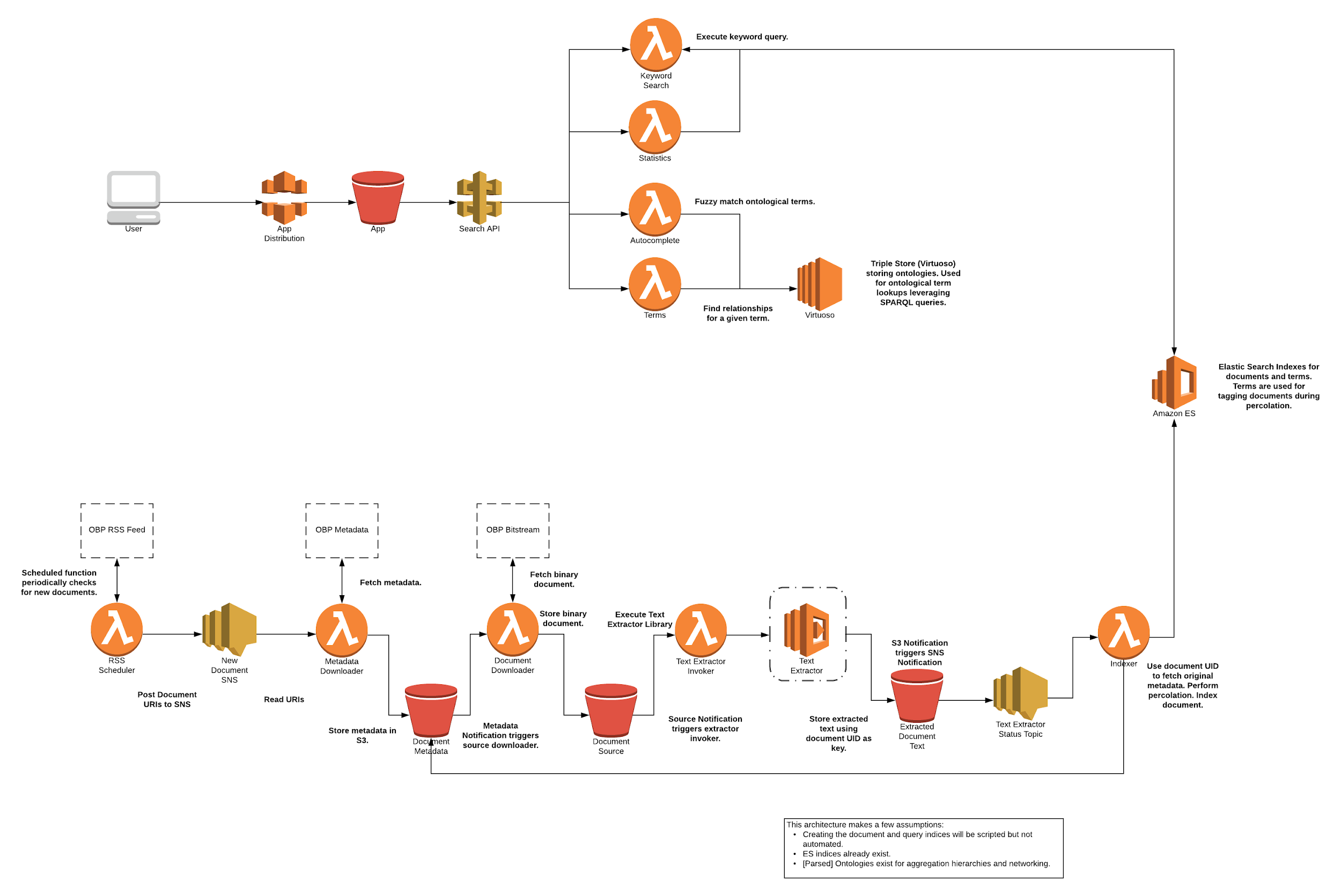
There are two main branches in the git repository representing the two main environments (production and UAT): *master* and *dev* branches. The *master* branch, to the best of our ability, is kept at an “always deployable” state. This means that if the situation arose where production needed to be re-deployed the *master* branch is always representative of the current production state.

## Project Infrastructure and Data

Project data consists of multiple Elasticsearch indices - document metadata and ontological tags - in addition to the ontologies and vocabularies supported by our system. Document metadata is provided by the DSpace API and is extracted and transformed by our system during ingest. Ontologies and vocabularies are managed by an instance of Virtuoso running on an AWS EC2 instance backed by an AWS EBS volume for storage. Since the OceanBestPractices (.org) portal is ***not*** the source of truth - meaning we do not collect any metadata or binary data that is not already maintained by DSpace - we do not actively backup any data stored within our system.

The infrastructure of our system is defined using multiple (modularized) AWS CloudFormation templates. These templates automate the process of creating an environment which allows for quick deployments. In the event of a system failure a new environment can be deployed within an hour. Once a new environment is deployed, automated scripts trigger the re-indexing of document metadata and binary download - populating the Elasticsearch indices with document metadata and ontological terms. This entire process, if necessary, can create an environment identical to the previous one within 1 hour.

# Appendix 1: System Architecture



**FIGURE 1**