



IN12C-02: (Polar) Domain Discovery with Sparkler

Siri-Jodha S Khalsa¹, Ruth Duerr², Chris A Mattmann³,
Nithin Krishna Ottilingam⁴, Karajeet Singh⁴, Luis Alberto Lopez¹, Omid Davtalab⁴, Simin Karvigh⁴



4

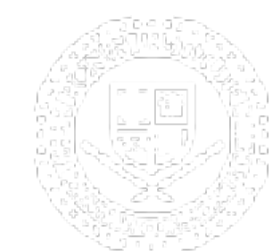


1



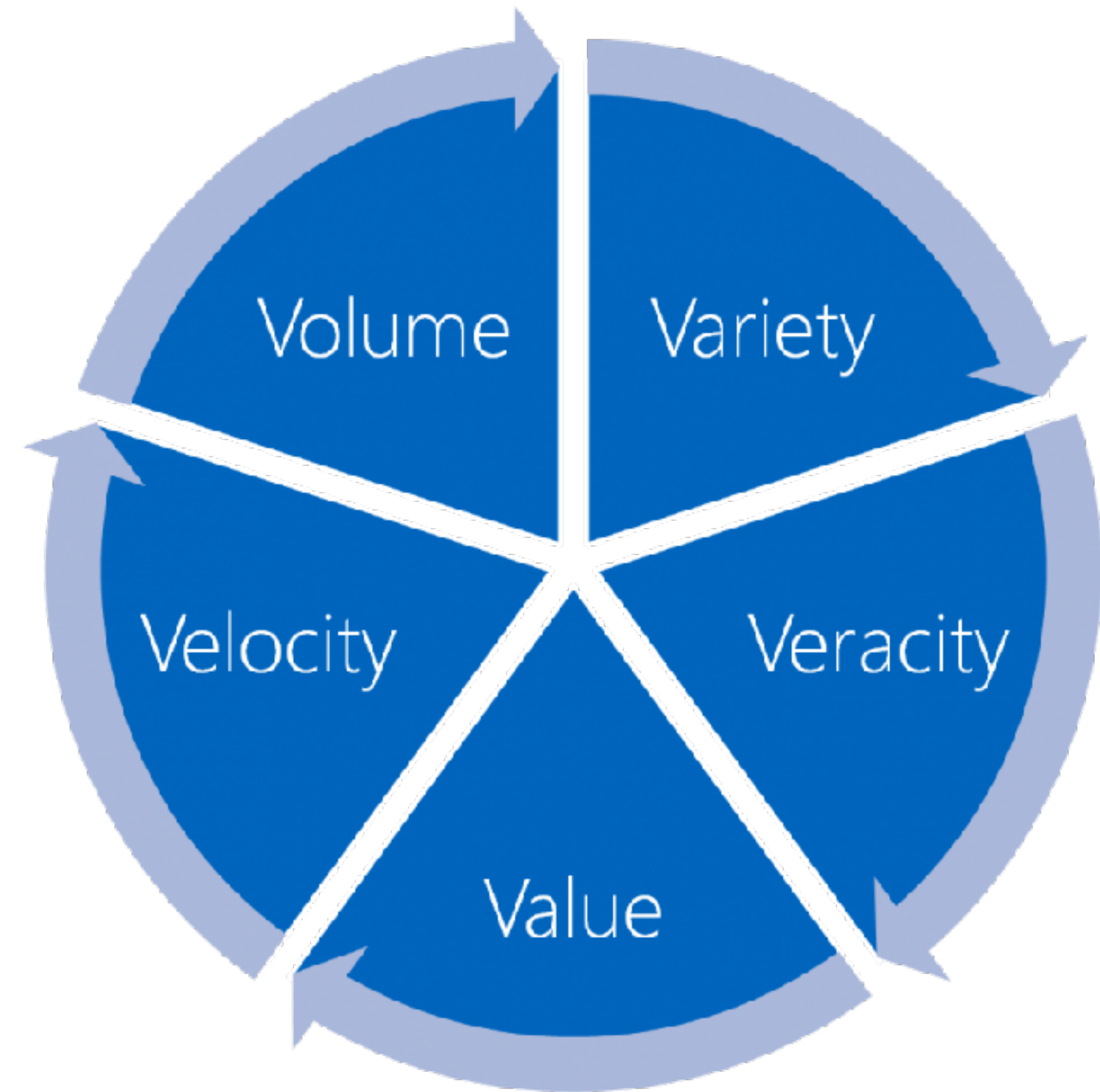
National Snow and Ice Data Center
Advancing knowledge of Earth's frozen regions

3



So, What's the Problem?

- Domain (polar) data is highly distributed
- Domain data is extremely diverse
- Cataloging all of it is an impossible task
- What if we just leave everything where it is and find it, as needed, through *focused crawling*?



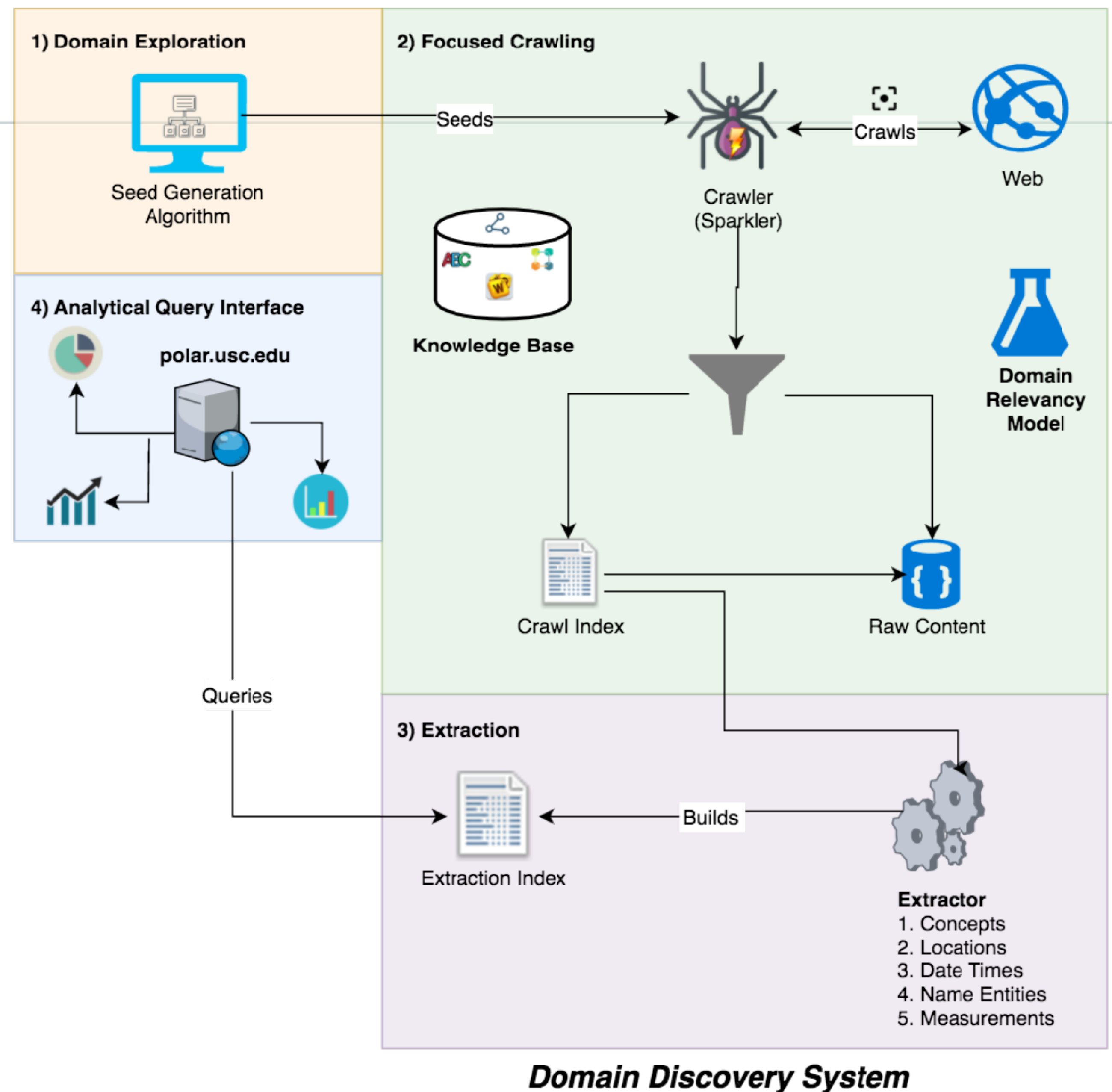
Applying “Big Data” Technology to Domain (here Polar) Data

- Make it possible to query the body of accumulated knowledge about a domain, using natural language and deep learning
 - Find the applicable data and documents
 - Evaluate the structure and contents to effectively extract information
 - Store and index the information
 - Create interface to query the content (using NLP/ML)

Polar Deep Insights Architecture

Leverages prior work done under the DARPA MEMEX (<http://memex.jpl.nasa.gov/>), NSF Polar CyberInfrastructure activities, and community workshops

1. Domain Exploration - Create a URL seed list and domain relevancy model
2. Focused Crawling - Crawl the web using the seed list and model
3. Extraction - Use a number of extractors to extract content from the documents returned by the crawl
4. Analytical Query Interface - Use a variety of analytical tools to explore the extracted content



Domain Exploration - Create a URL seed list and domain relevancy model

We are currently exploring two paths:

1. Subject Matter Expert model generation (from knowledge base, glossaries, search terms, and seed URL's)
2. Semi-automated model generation

Intent to compare efficiency and accuracy of each path (and their variations)

Two test cases so far:

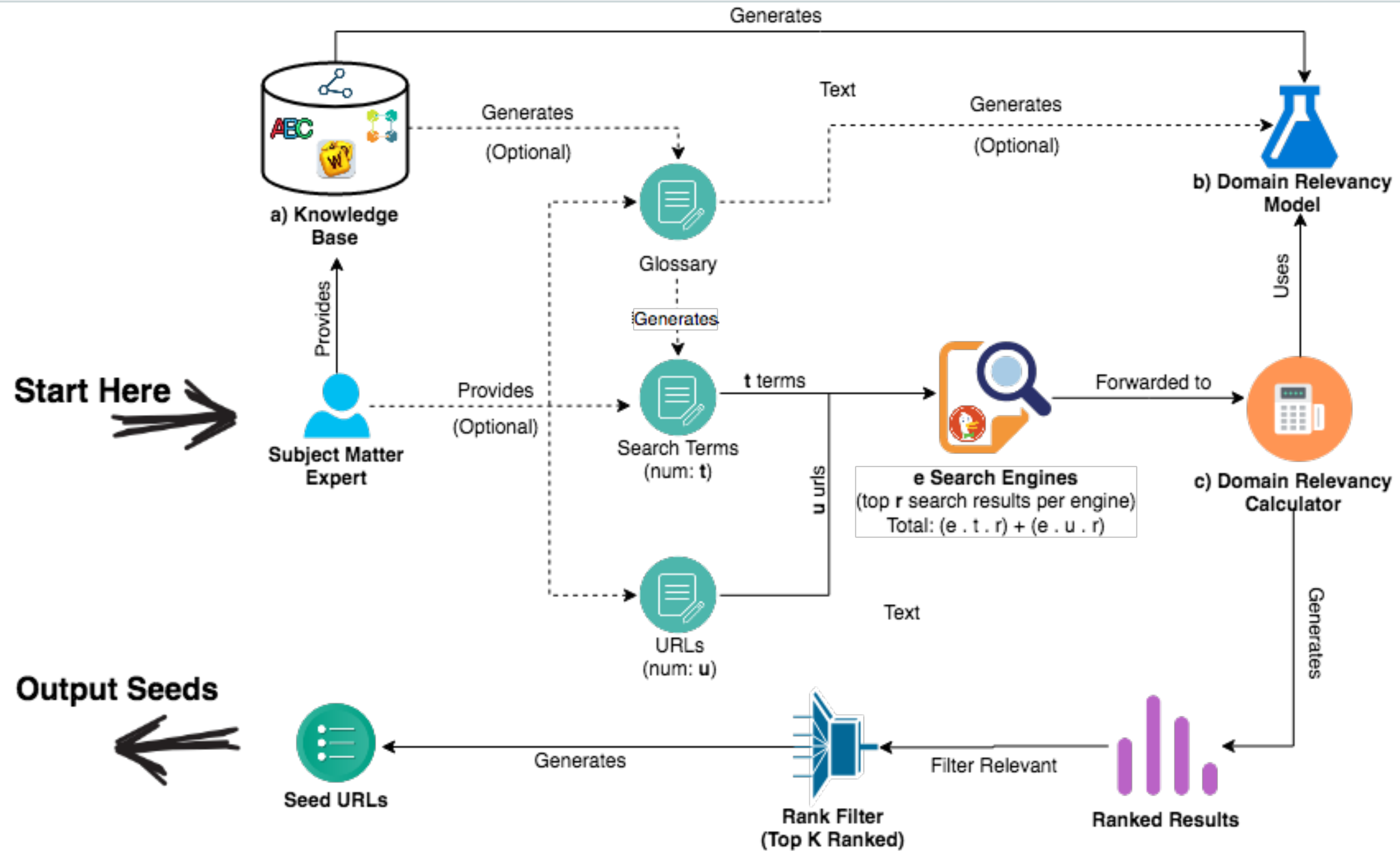
1. Sea ice based off the NSF-funded SSIII project's sea ice ontologies
2. Search terms, seed URL's, etc. provided by Jay Pearlman and Pier Luigi Buttigieg for their Integrated Oceanographic Data and Information Exchange Best Practices work

1) Domain Exploration



Seed Generation
Algorithm

Domain Exploration - Subject Matter Expert Model Generation



Domain Exploration - Semi-automated Model Generation

- Finding relevant URLs to crawl
 - Start with small set of seeds provided by domain experts
 - Feed these to general search engines and rank the additional found links according to text similarity and other measures
- Domain experts rate these URLs for relevancy
 - This annotated set of URLs then used to train a machine learning model to predict the 'domain relevance' of a given document



Domain Exploration - Semi-automated Model Generation

Domain Discovery - Seed Generation Update Model

1 Generate a Model

Minimum 10 each
78 66 108

[More Options »](#)

2 Create a Seed File
[Import Seed File](#)

3 Start the Crawl
[Start Crawler](#)
[More Options »](#)

4 Visit the Crawl Dashboard

Title: Best Practices for Website Navigati
URL: <https://ocean19.com/blog/best-pract>

```
<iframe src="https://www.googletagmanager.com/ns.html?id=GTM-NG85BQL" height="0" width="0" style="display:none;visibility:hidden"></iframe>
```

Title: Best Practices Teaser - Data.gov
URL: <https://www.data.gov/ocean/best-pra>

Title: OCADS - Guide to Best Practices for
URL: <https://www.nodc.noaa.gov/ocads/cce>

Title: Ocean - Best practices, tips and fu
URL: <https://www.classy.org/blog/ocean/>

Title: Welcome to the Frontpage
URL: <http://www.oceandatastandards.org/>

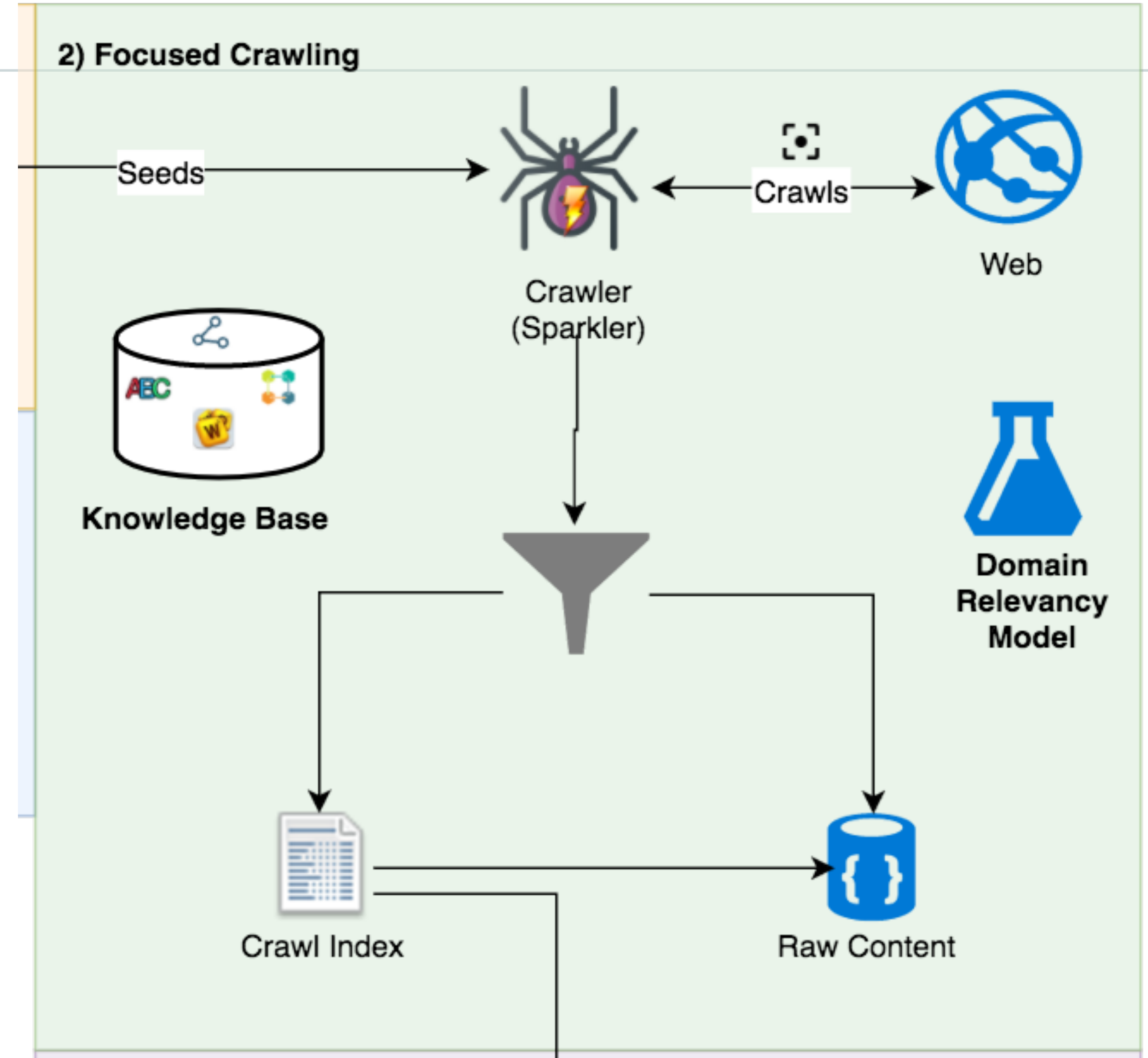
Title: Ocean - Best Practices - Data.gov
URL: <https://www.data.gov/ocean/best-pra>

IN12C-02: (Polar) Domain Discovery with Sparkler, presented by Ruth Duerr at the 2017 AGU Fall Meeting

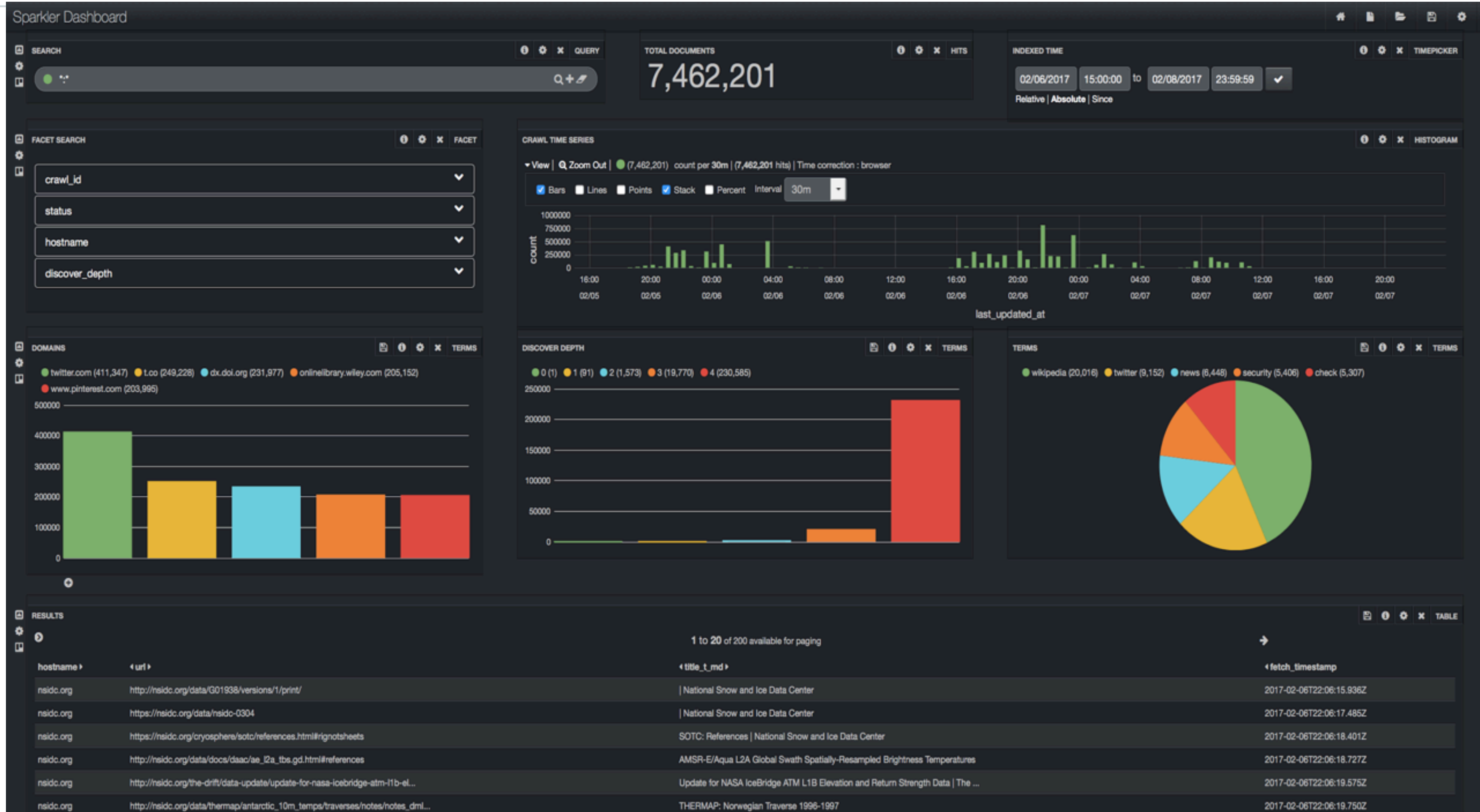
Ronin Institute

Focused Crawling

- Sparkler (<https://github.com/USCDataScience/sparkler>) is an extensible, highly scalable Web crawler that runs on top of Spark (vice Hadoop)
- Uses the domain relevancy model to find resources
- Avoids disrupting hosts being crawled
 - Partitions URLs by hostname and every node gets a different host to crawl
 - Inserts time delays between successive requests

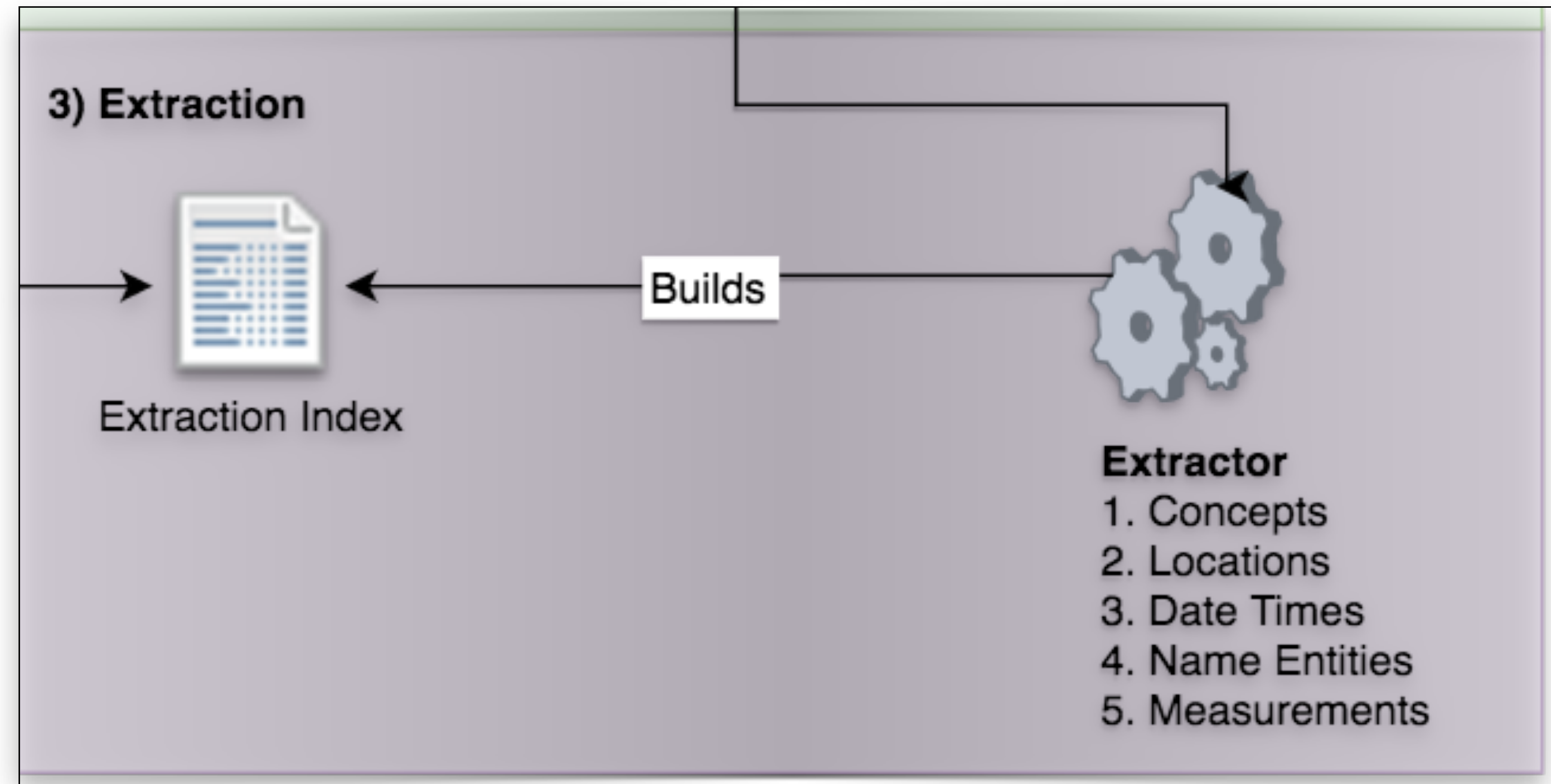


Focused Crawling Dashboard



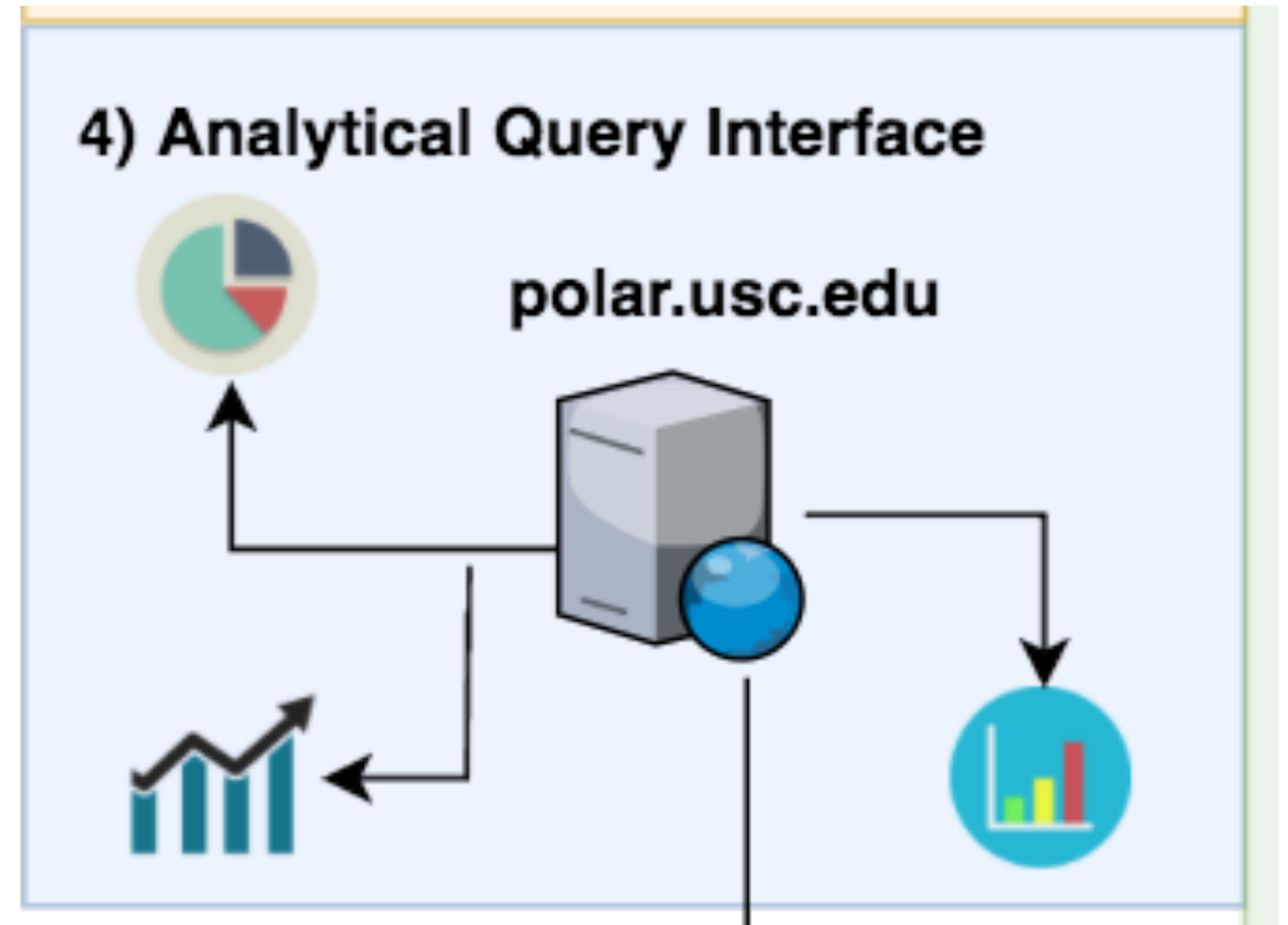
Extraction

- Detects and extracts metadata, text, URLs
- Toolkit of parsers to extract
 - Concepts
 - Geographic locations
 - Dates and Times
 - Named Entities
 - Numerical measurements
- Creates an index for the extracted content

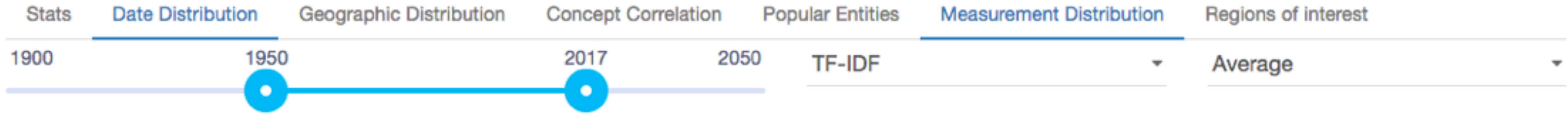


Analytical Query Interfaces

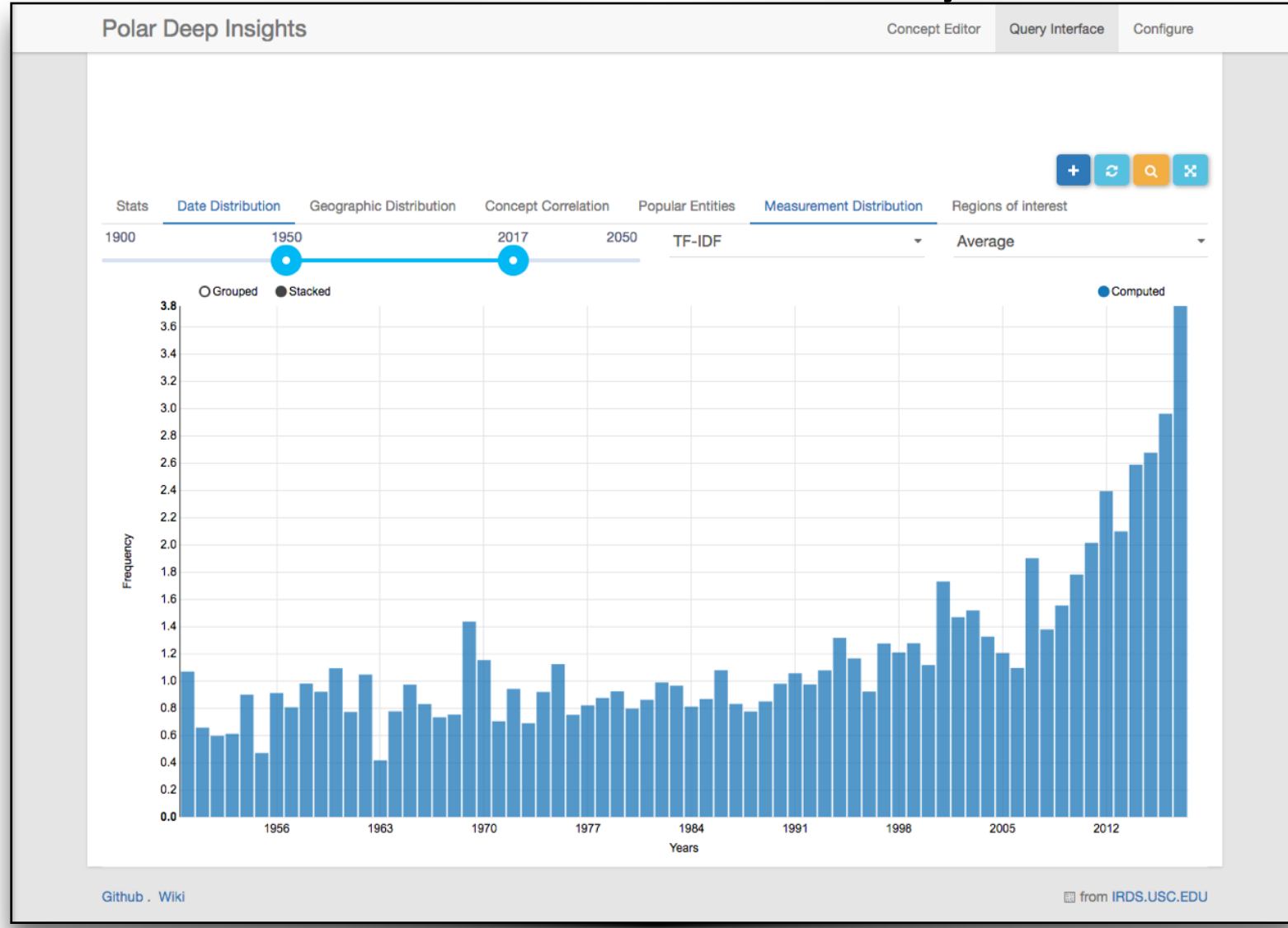
- This vast store of information is of little use without an efficient and intuitive means of querying it
- Polar Data Insights is experimenting with various tools that an user can interact with through different dashboards to query and visualize the data
 - Banana
 - FacetView



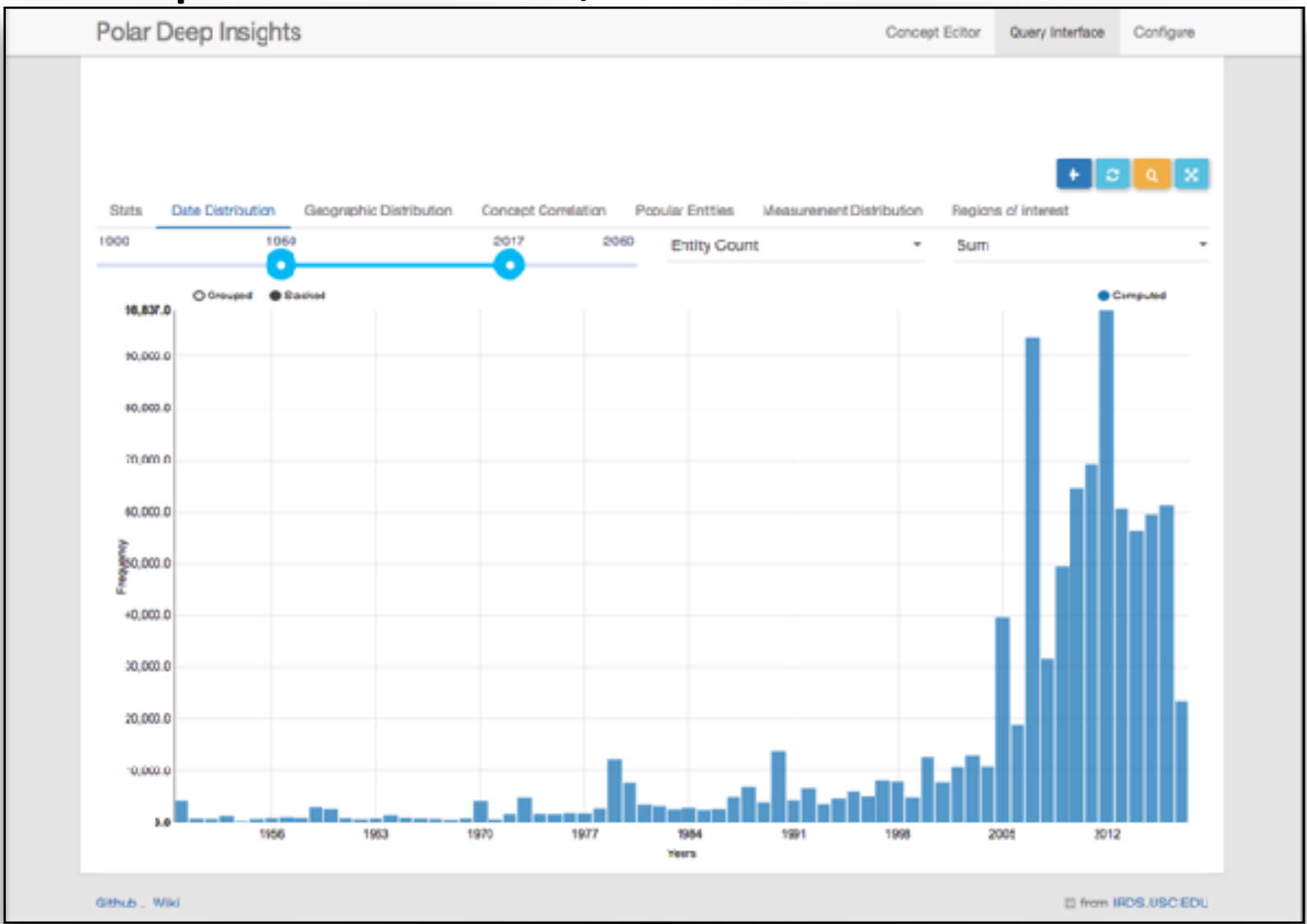
Polar Deep Insights - Banana based query and analysis



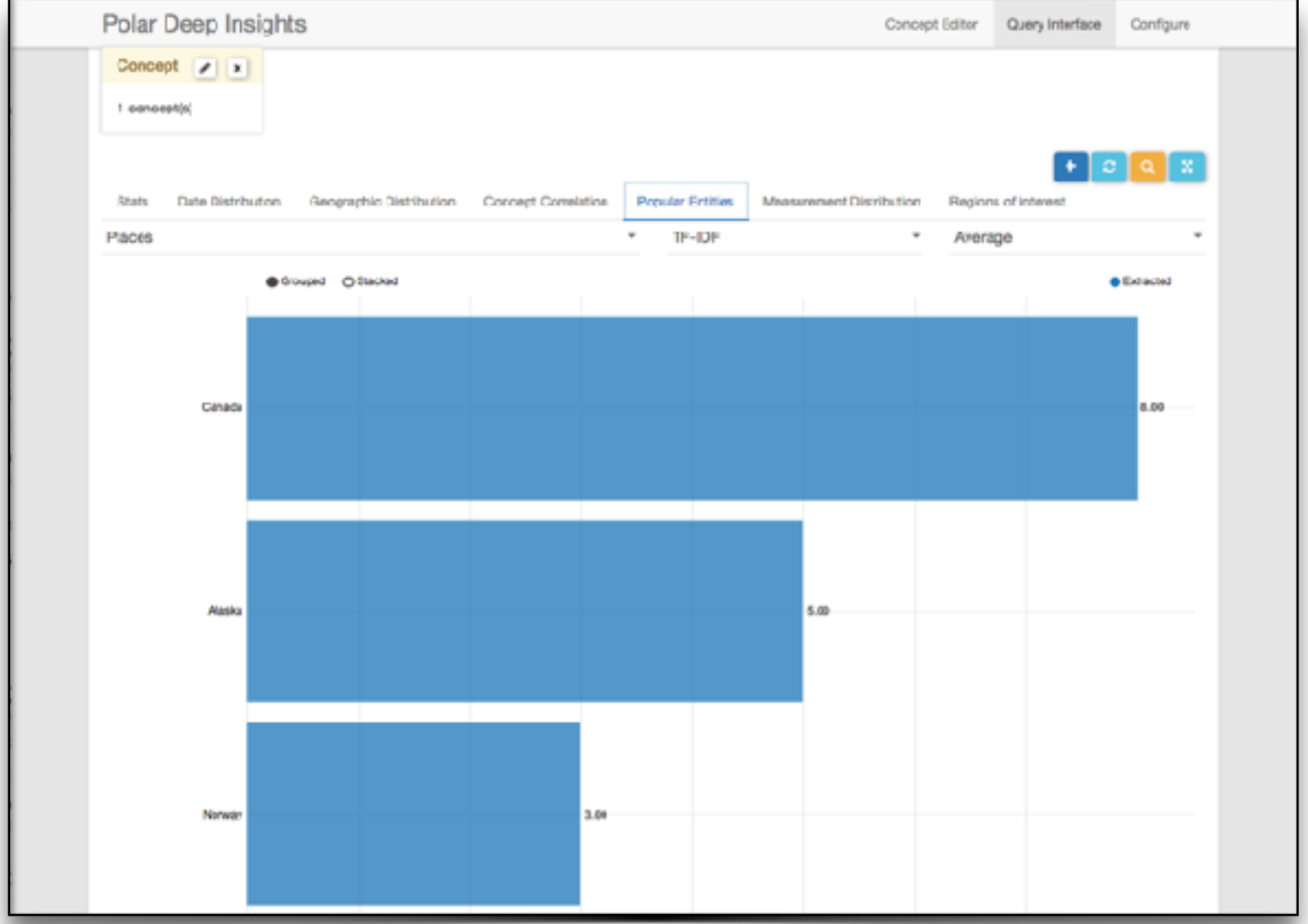
Distribution of documents by date



Distribution of documents that mention icebergs again by date - What's up with the spikes in 2005, 2007 and 2012?

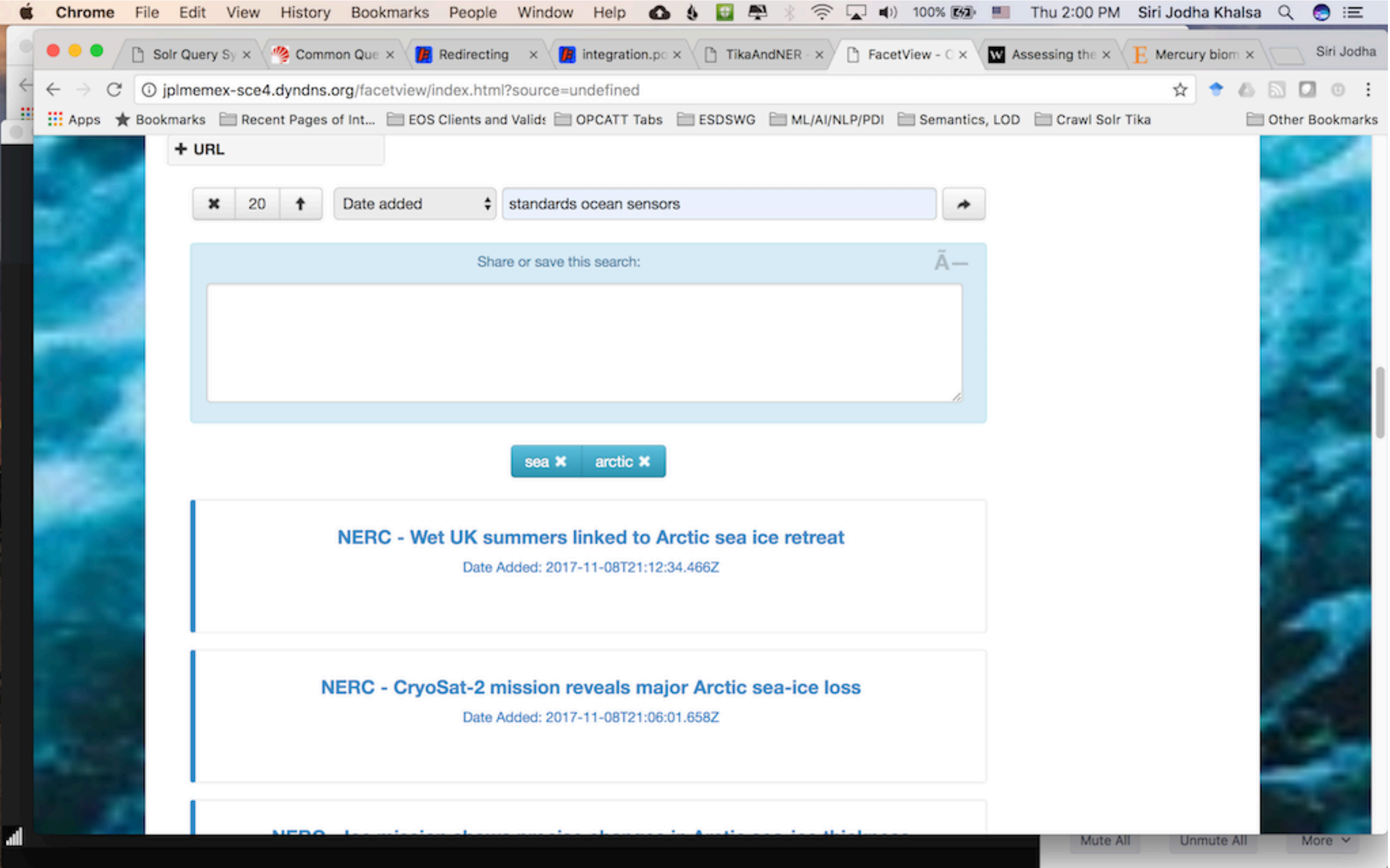


Locations mentioned in documents that mentioned icebergs and that something was ice-bound.



Polar Deep Insights - Facet-view based query and analysis

Query for documents mentioning the words “standards”, “ocean” and “sensors” using a keyword facet with terms “arctic” and “sea”



Polar Deep Insights - Facet-view based query and analysis

Query for documents mentioning the words “standards”, “ocean” and “sensors” using a keyword facet with terms “arctic” and “sea”

The screenshot shows a web browser window with a search results page. The browser's address bar shows the URL: `jplmemex-sce4.dyndns.org/facetview/index.html?source=undefined`. The page displays three search results, each in a white box with a blue border. The first result is titled "Frost flowers growing in the Arctic ocean-atmosphere-sea ice-snow interface: 1. Chemical composition - Douglas - 2012 - Journal of Geophysical Research: Atmospheres - Wiley Online Library" and has a date added of 2017-10-30T17:21:23.554Z. The second result is titled "Assessing the potential impacts of declining Arctic sea ice cover on the photochemical degradation of dissolved organic matter in the Chukchi and Beaufort Seas - Logvinova - 2015 - Journal of Geophysical Research: Biogeosciences - Wiley Online Library" and has a date added of 2017-11-08T00:00:06.327Z. The third result is titled "Annual cycles of pCO2sw in the southeastern Beaufort Sea: New understandings of air-sea CO2 exchange in arctic polynya regions - Else - 2012 - Journal of Geophysical Research: Oceans - Wiley Online Library" and has a date added of 2017-11-09T11:52:33.264Z. At the bottom of the page, there is a pagination control showing "1 - 20 of 21" and a "next" button. The browser's taskbar at the bottom shows "Mute All", "Unmute All", and "More" options.

Polar Deep Insights - Facet-view based query and analysis

Query for documents mentioning the words “standards”, “ocean” and “sensors” using a keyword facet with terms “arctic” and “sea”

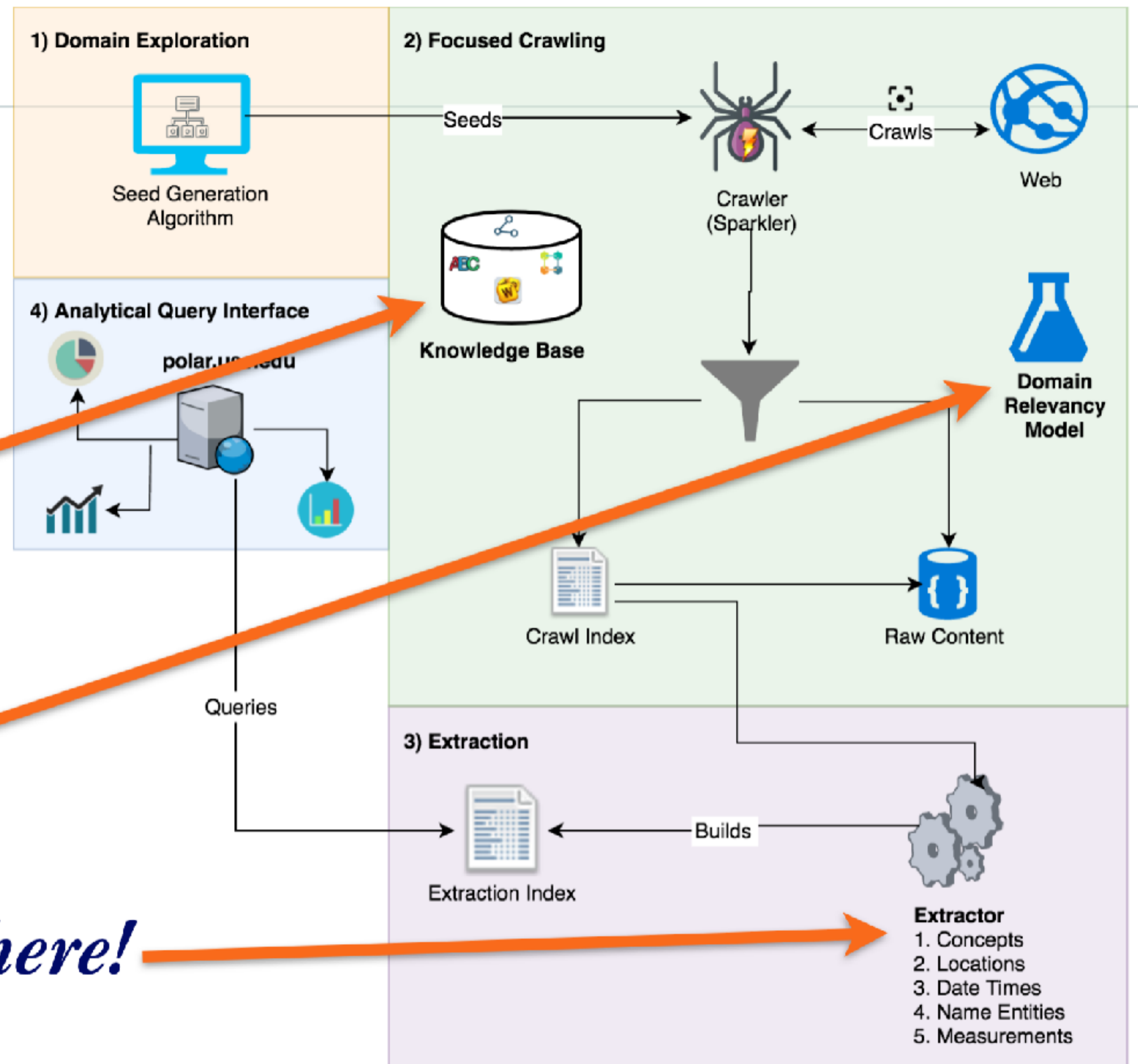
The image displays three overlapping browser windows. The leftmost window shows a search interface with a 'URL' field and a search button. The middle window shows a search results page with a list of documents, including 'Frost flowers g...', 'Chemical co...', 'Assessing photochemi...', 'Beaufort...', and 'Annual cycles of air-sea C...'. The rightmost window shows a full document page from 'onlinelibrary.wiley.com' with the URL 'doi/10.1002/2015JG003052/full'. The document text includes a paragraph about model fitting and validation, a section titled '2.5 Dissolved Organic Carbon Analysis' with a highlighted sentence: 'Standards were prepared by the volumetric dilution of a stock solution containing 500 μM DOC (potassium hydrogen phthalate, analytical grade) to produce the following series of standards: 0, 2, 5, 8, 10, 25, 50, 75, and 100 μM DOC.', and a section titled '3 Results' with a subsection '3.1 Photodegradation of CDOM'. A sidebar on the right of the document page contains a table of contents with items like 'Abstract', '1 Introduction', '2 Methods', '3 Results', '4 Discussion', '5 Implications and Conclusions', 'Acknowledgments', 'References', 'Related Content', and 'Citing Literature'. A search bar at the top right of the document page shows the term 'standards' with a count of '2/4'.

Semantics is Everywhere!

Semantics is here!

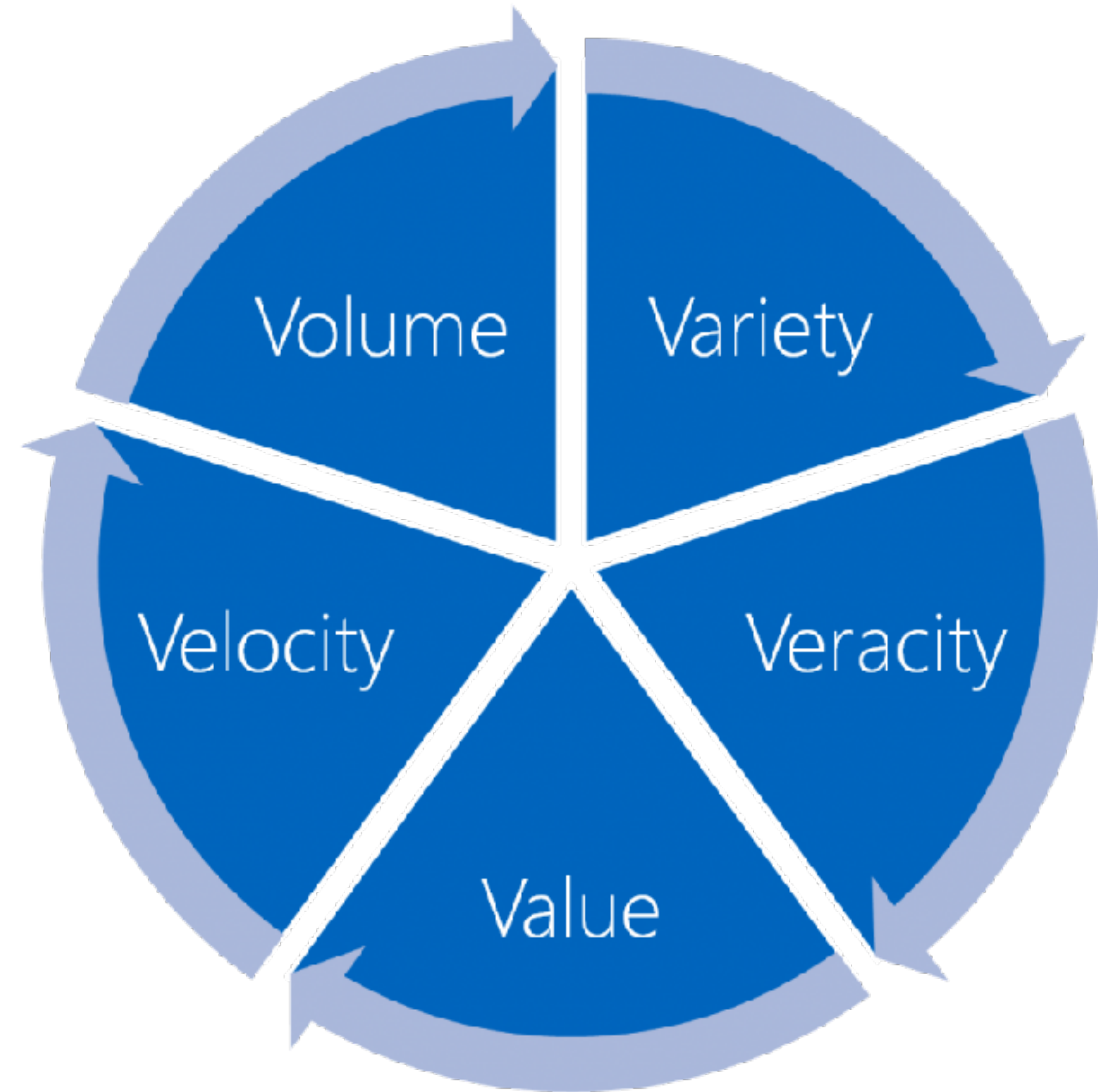
And here!

And here!



Summary and Conclusions

- A new approach to data discovery and information extraction is required to make effective use of the wealth of textual and scientific data that is being generated
- An Open Source framework fosters community involvement in the development, and responsive evolution of the necessary tools
- These tools can provide the ability to address grand challenge questions concerning the state and trajectory of the Earth System and its Polar regions



Acknowledgements

This work would not have been possible without funding by NSF through ICER grant #1639675

