2

IN12C-02: (Polar) Domain Discovery with Sparkler

Siri-Jodha S Khalsa[1], **Ruth Duerr**[2], Chris A Mattmann[3], Nithin Krishna Ottilingam[4], Karajeet Singh[4], Luis Alberto Lopez[1], Omid Davtalab[4], Simin Karvigh[4]

4

1

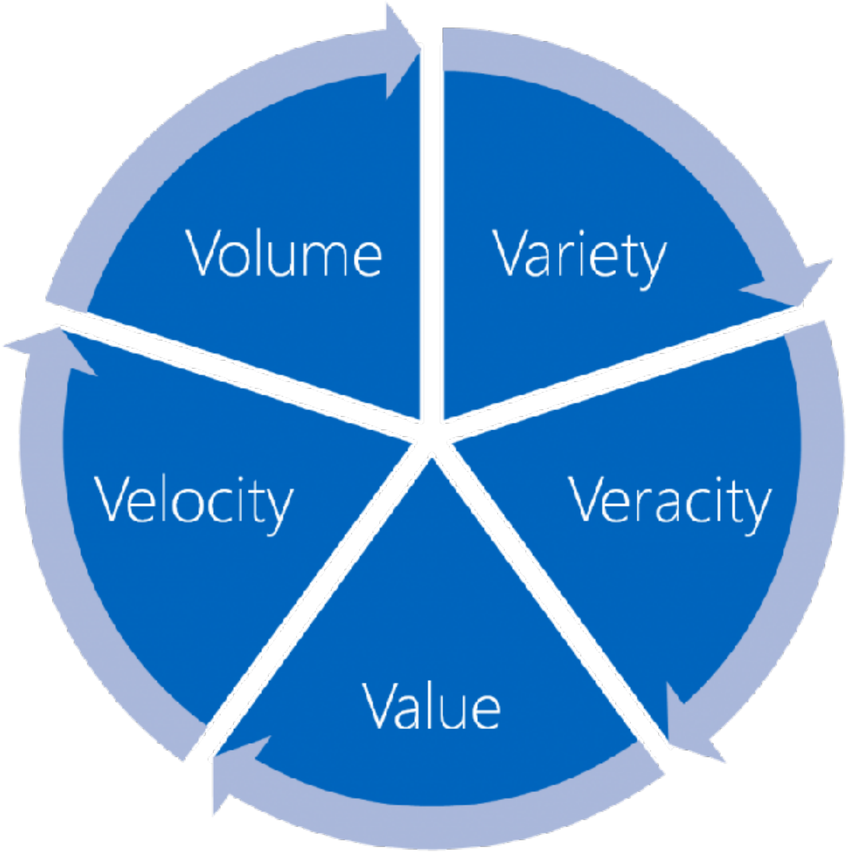
National Snow and Ice Data Center

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*?



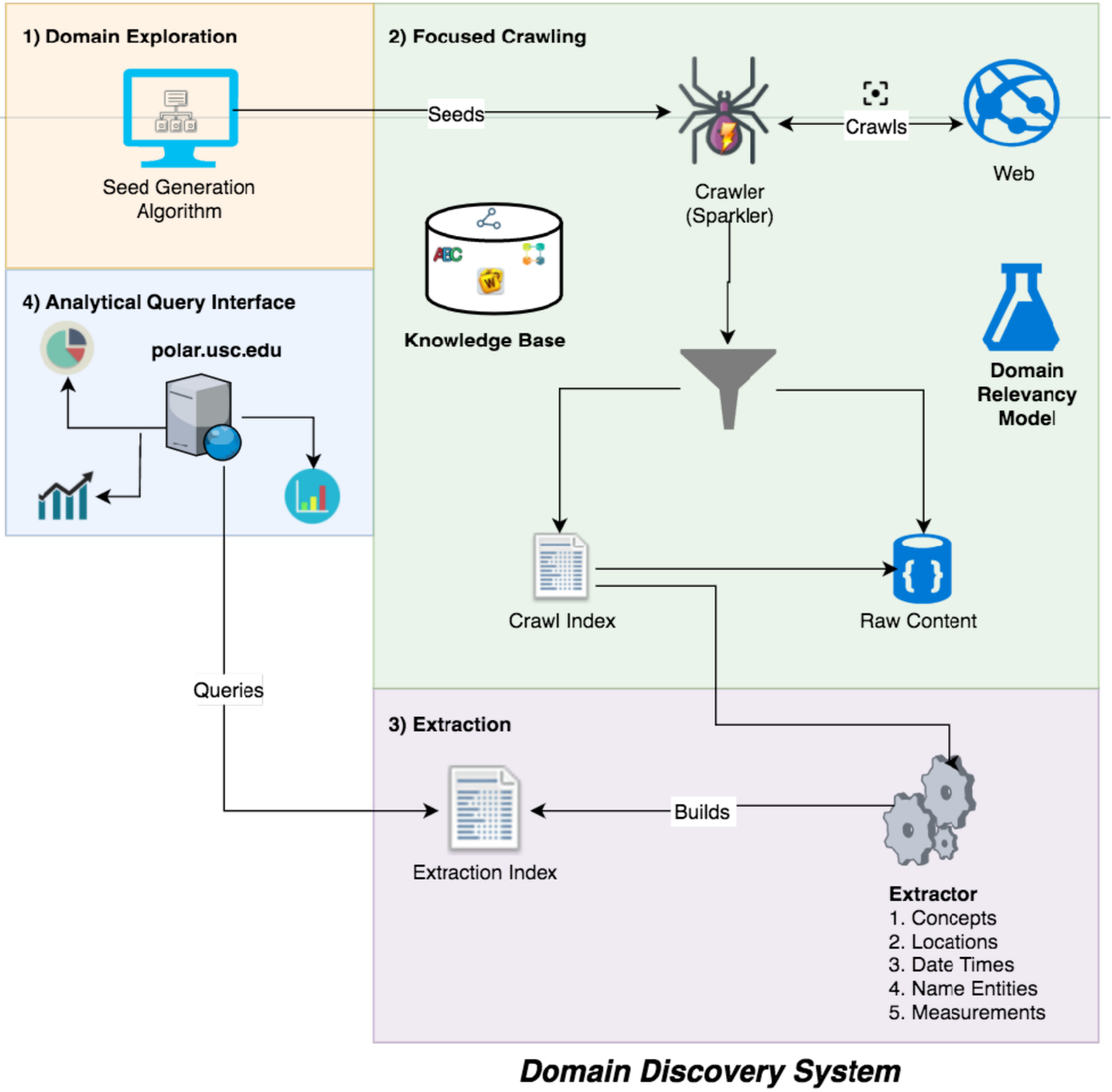
Applying "Big Data" Technology to Domain (here Polar) Data

- 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

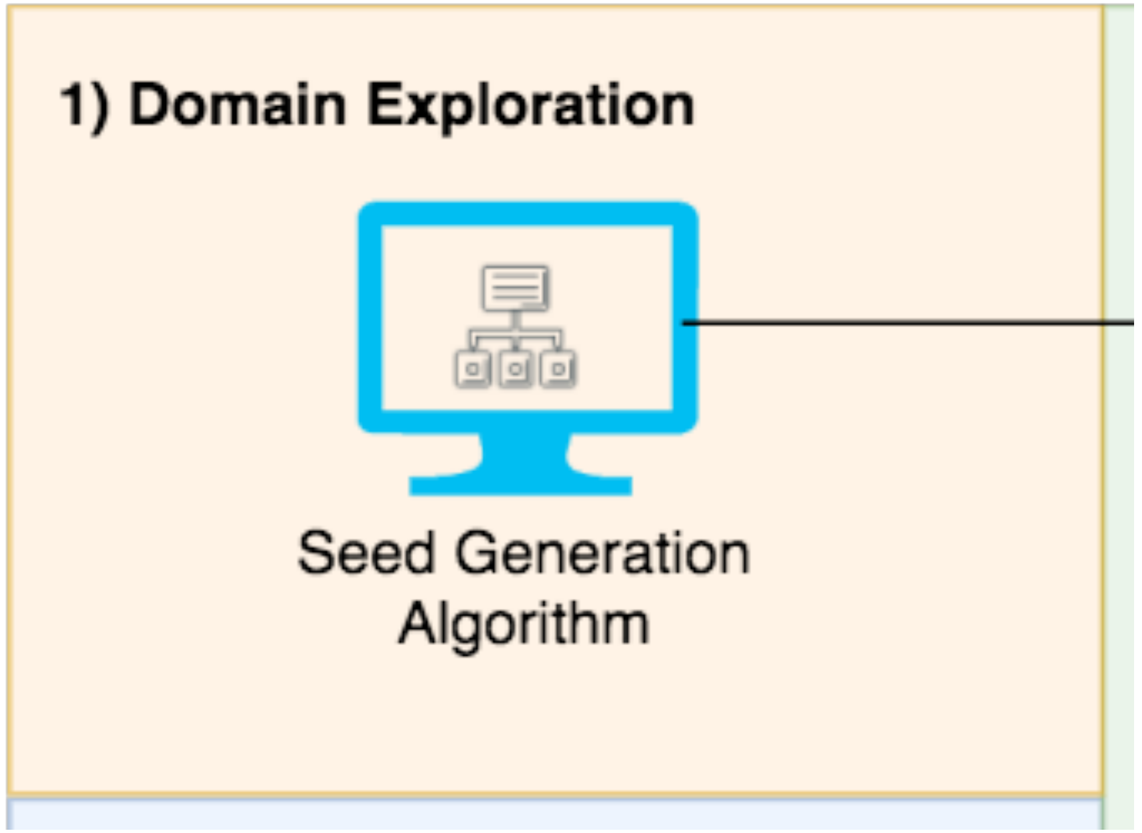
We are currently exploring two paths:

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

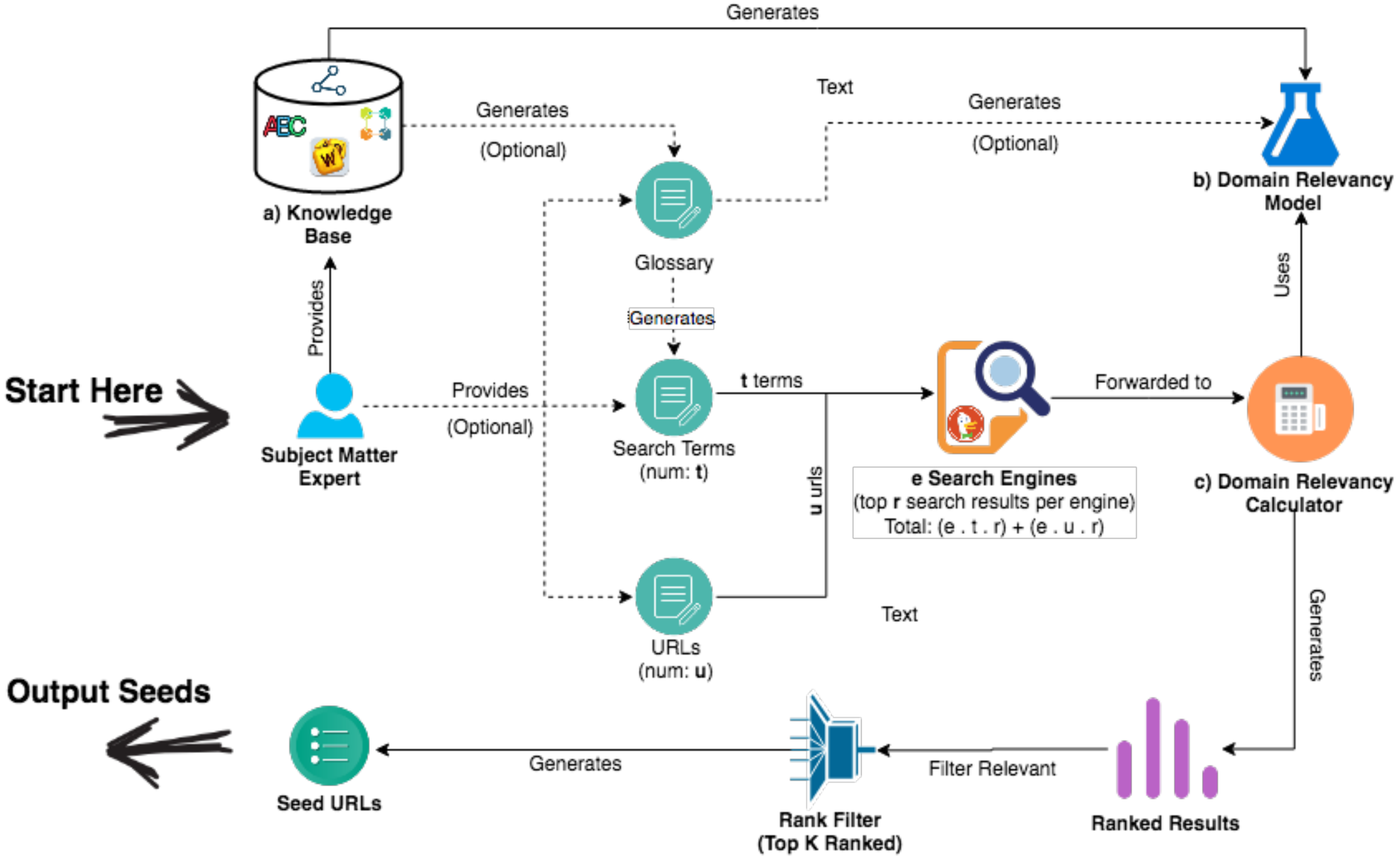
Intent to compare efficiency and accuracy of each path (and their variations)

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



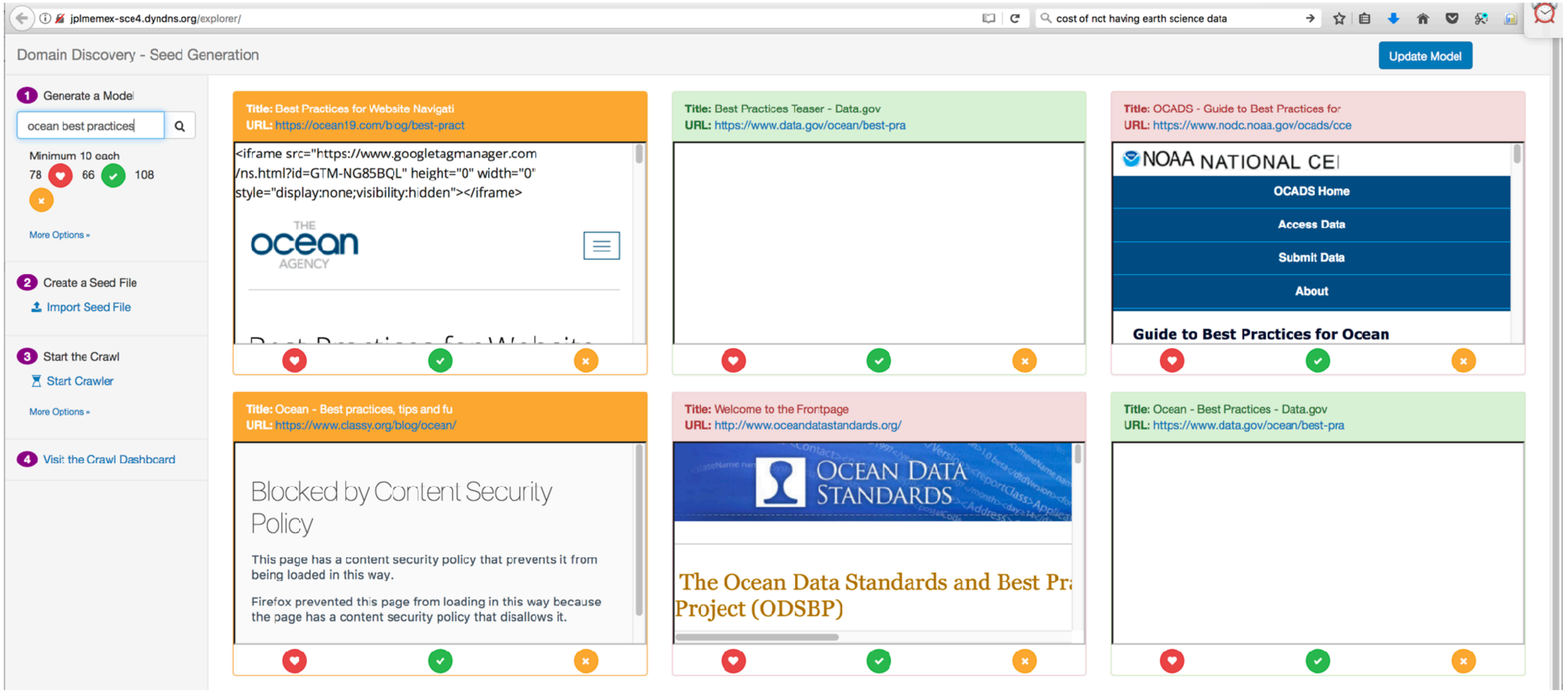
Domain Exploration - Subject Matter Expert Model Generation



Domain Exploration - Semi-automated Model Generation

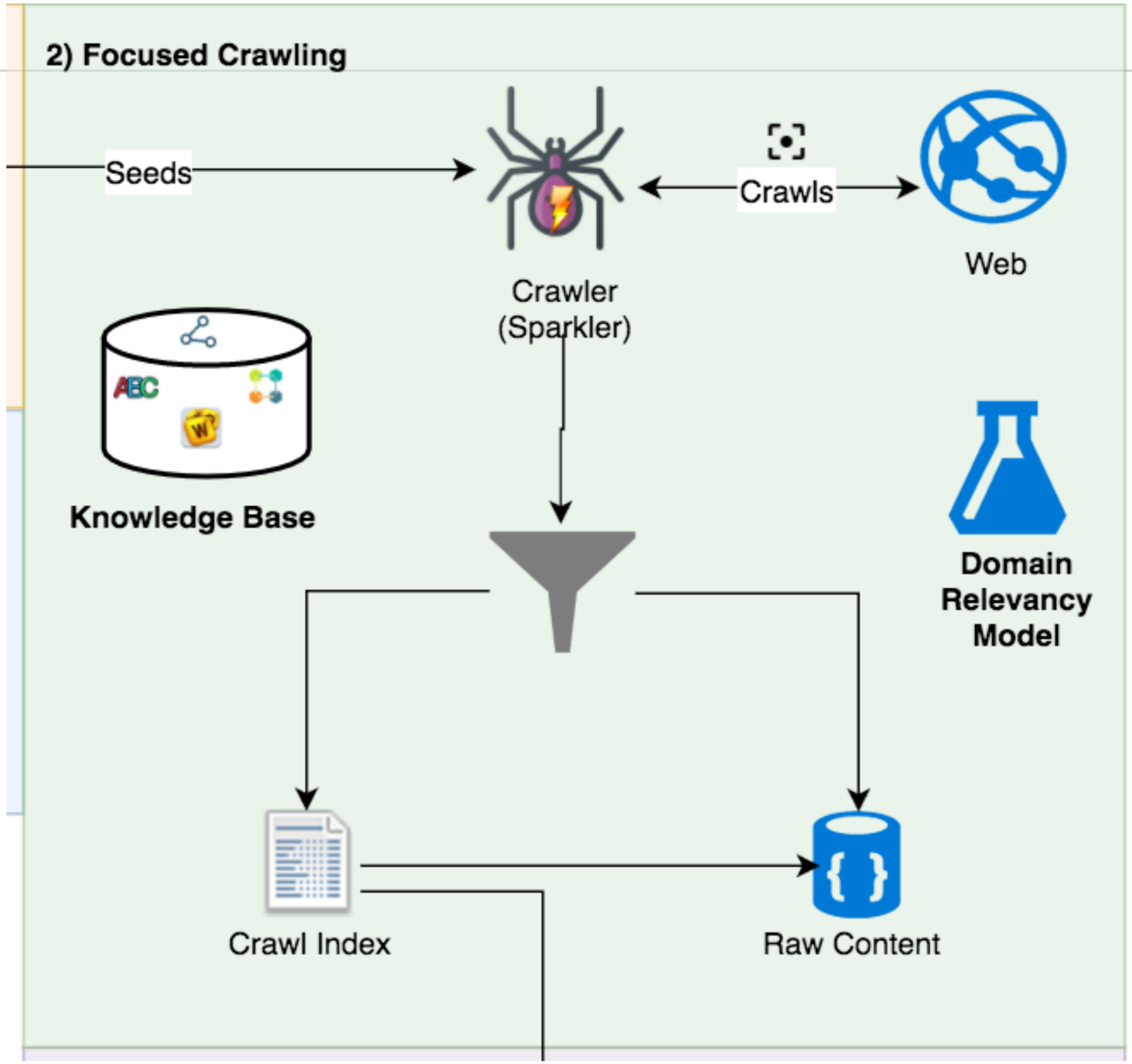
- Finding relevant URLs to crawl
  - Start with small set of seeds provided by domain experts
  - Feed these to general search engines and rank the additional found links according to text similarity and other measures
- Domain experts rate these URLs for relevancy
  - This annotated set of URLs then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

Domain Exploration - Semi-automated Model Generation

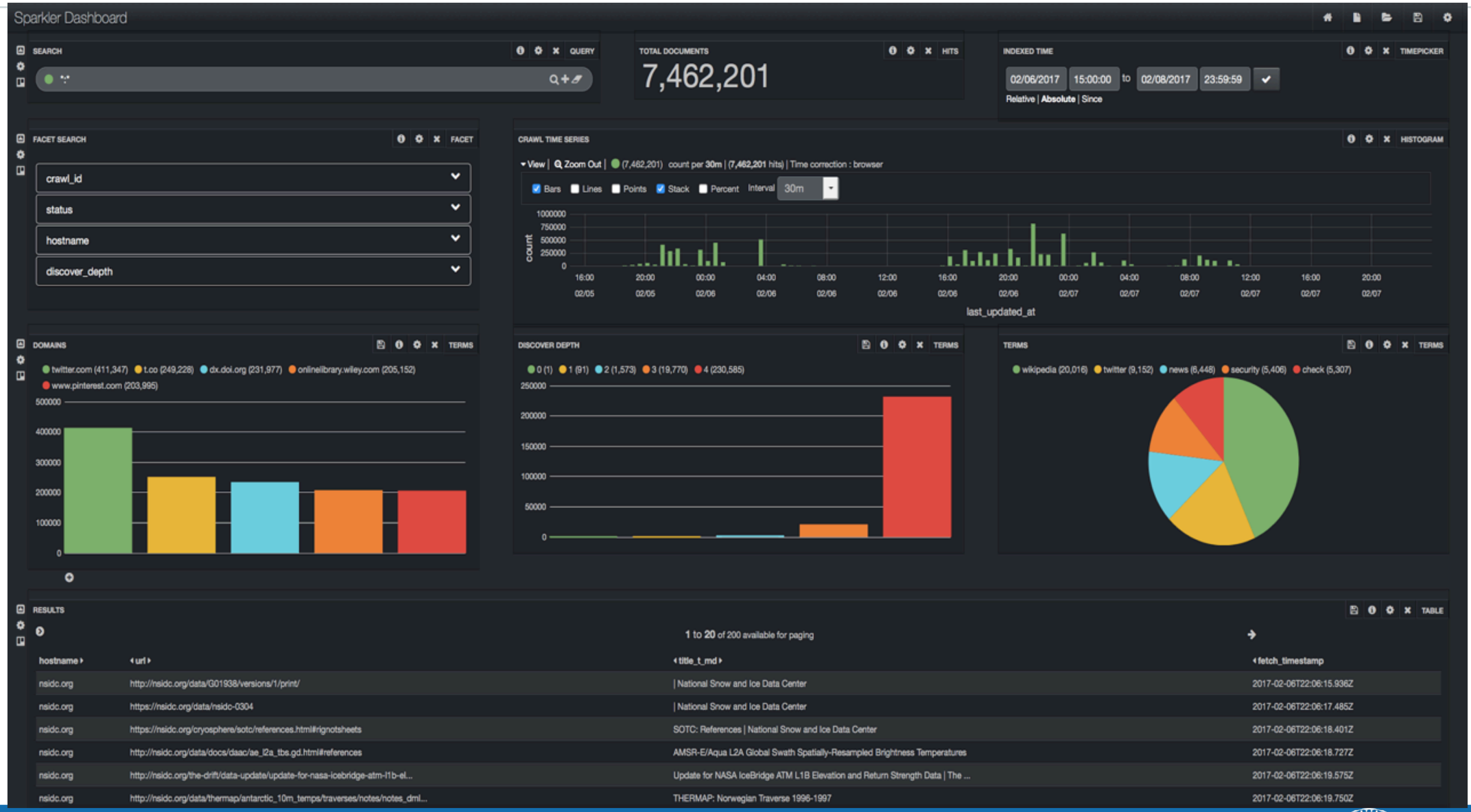


Focused Crawling

- 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

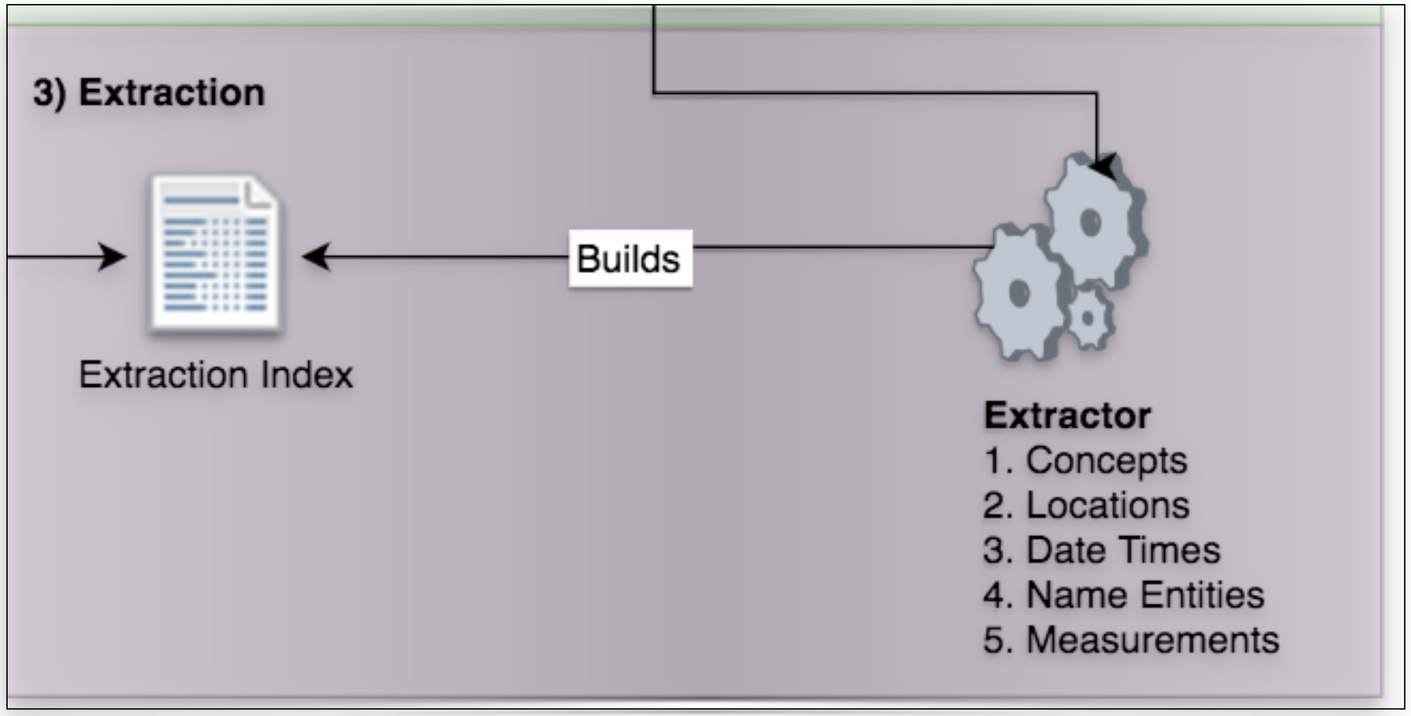


Focused Crawling Dashboard



Extraction

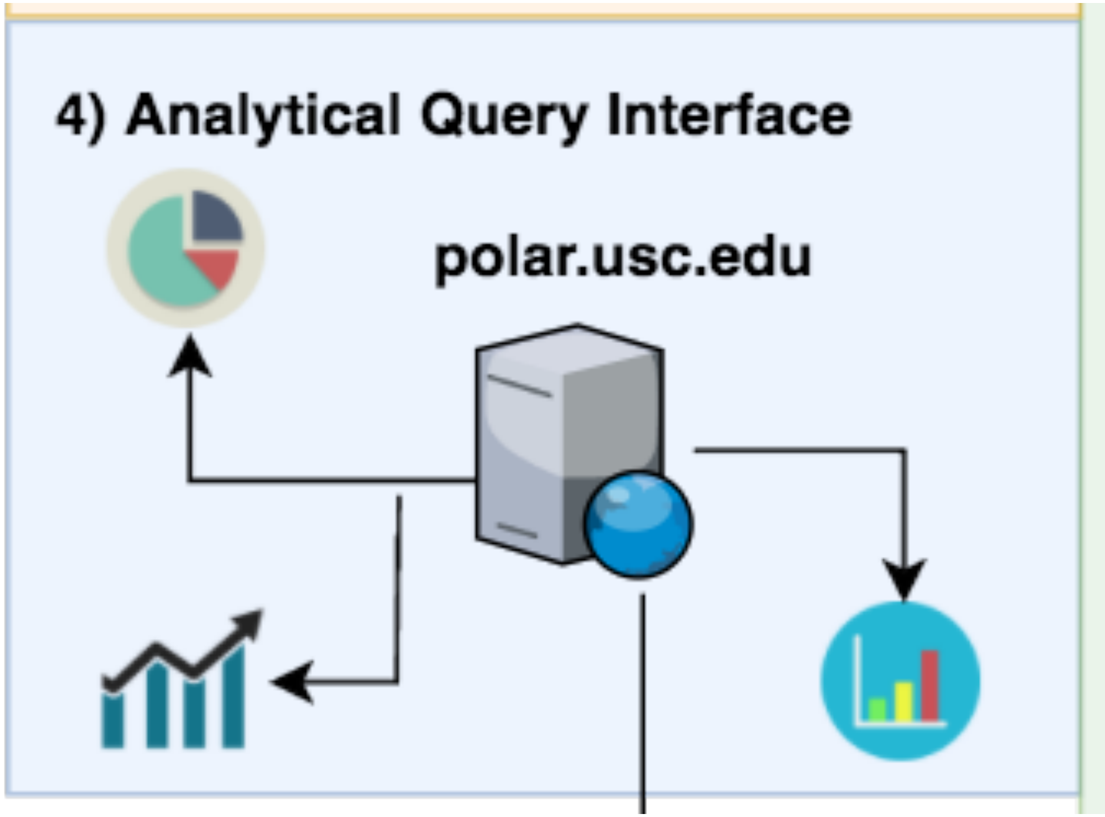
- Detects and extracts metadata, text, URLs
- Toolkit of parsers to extract
  - Concepts
  - Geographic locations
  - Dates and Times
  - Named Entities
  - Numerical measurements
- Creates an index for the extracted content



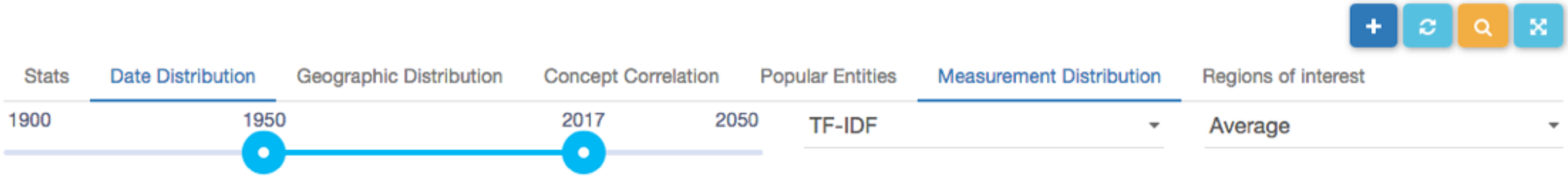
Apache™ Tika

Analytical Query Interfaces

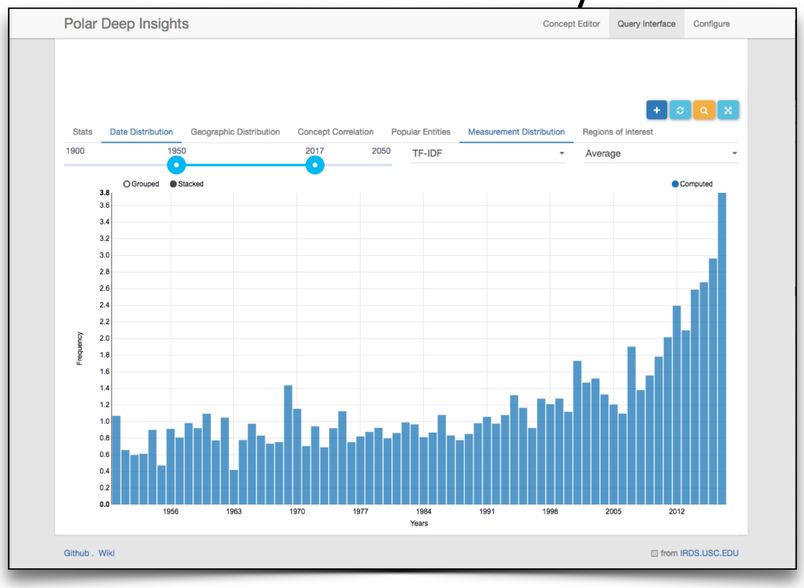
- This this vast store of information is of little use without an efficient and intuitive means of querying it
- Polar Data Insights is experimenting with various tools that an user can interact with through different dashboards to query and visualize the data
  - Banana
  - FacetView



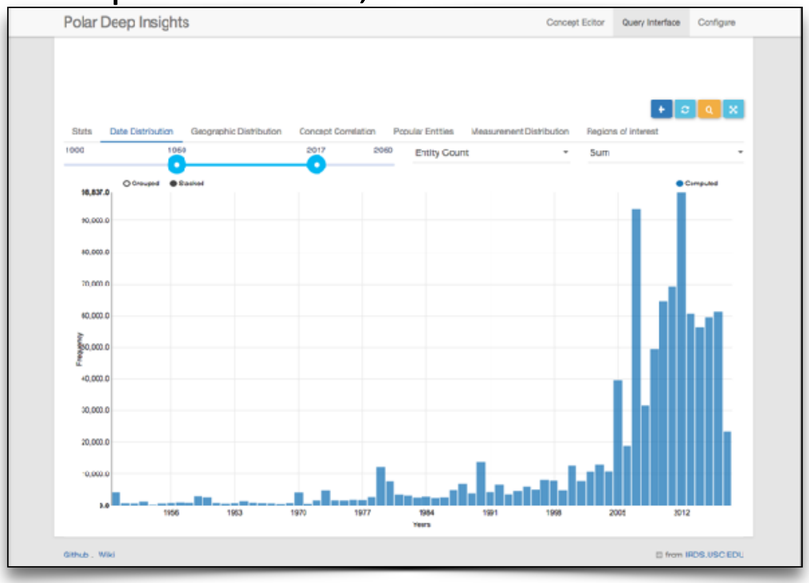
Polar Deep Insights - Banana based query and analysis



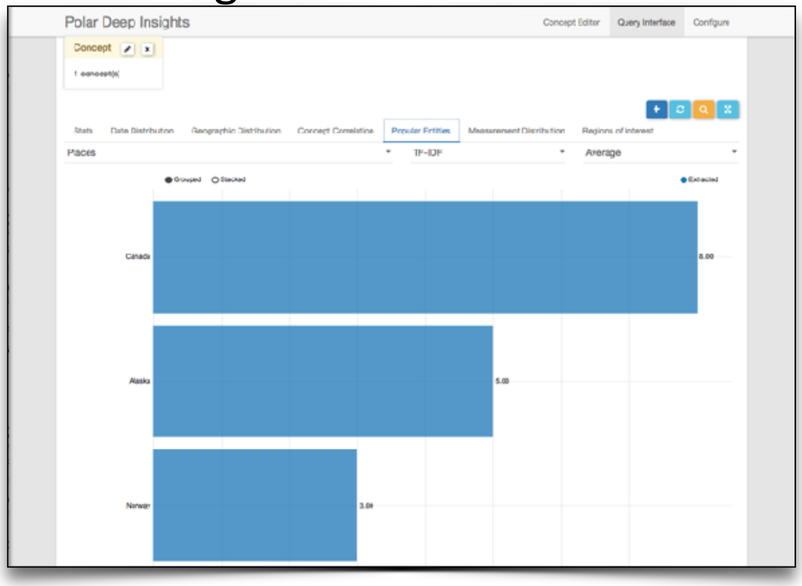
Distribution of documents by date



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

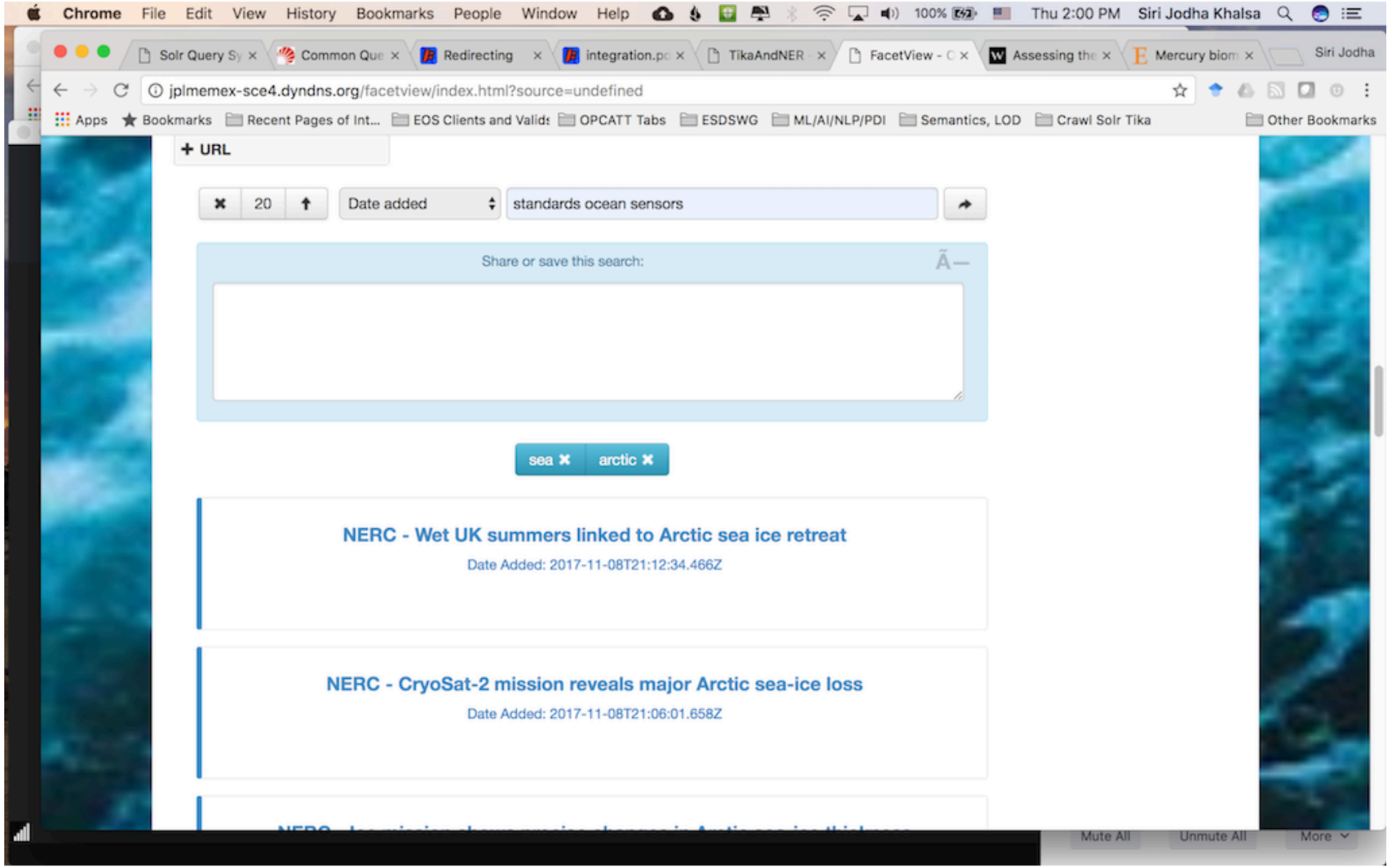


Locations mentioned in documents that mentioned icebergs and that something was ice-bound.



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”

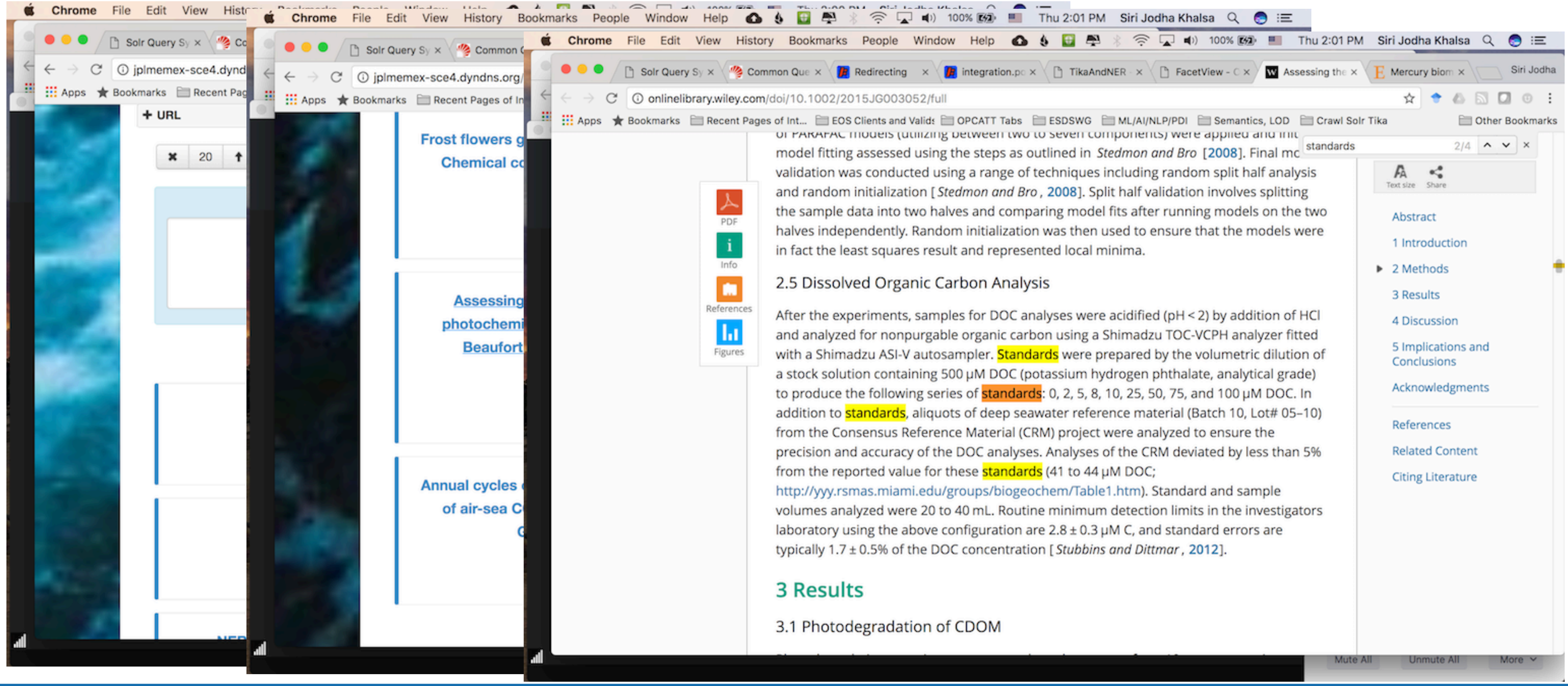


Polar Deep Insights - Facet-view based query and analysis

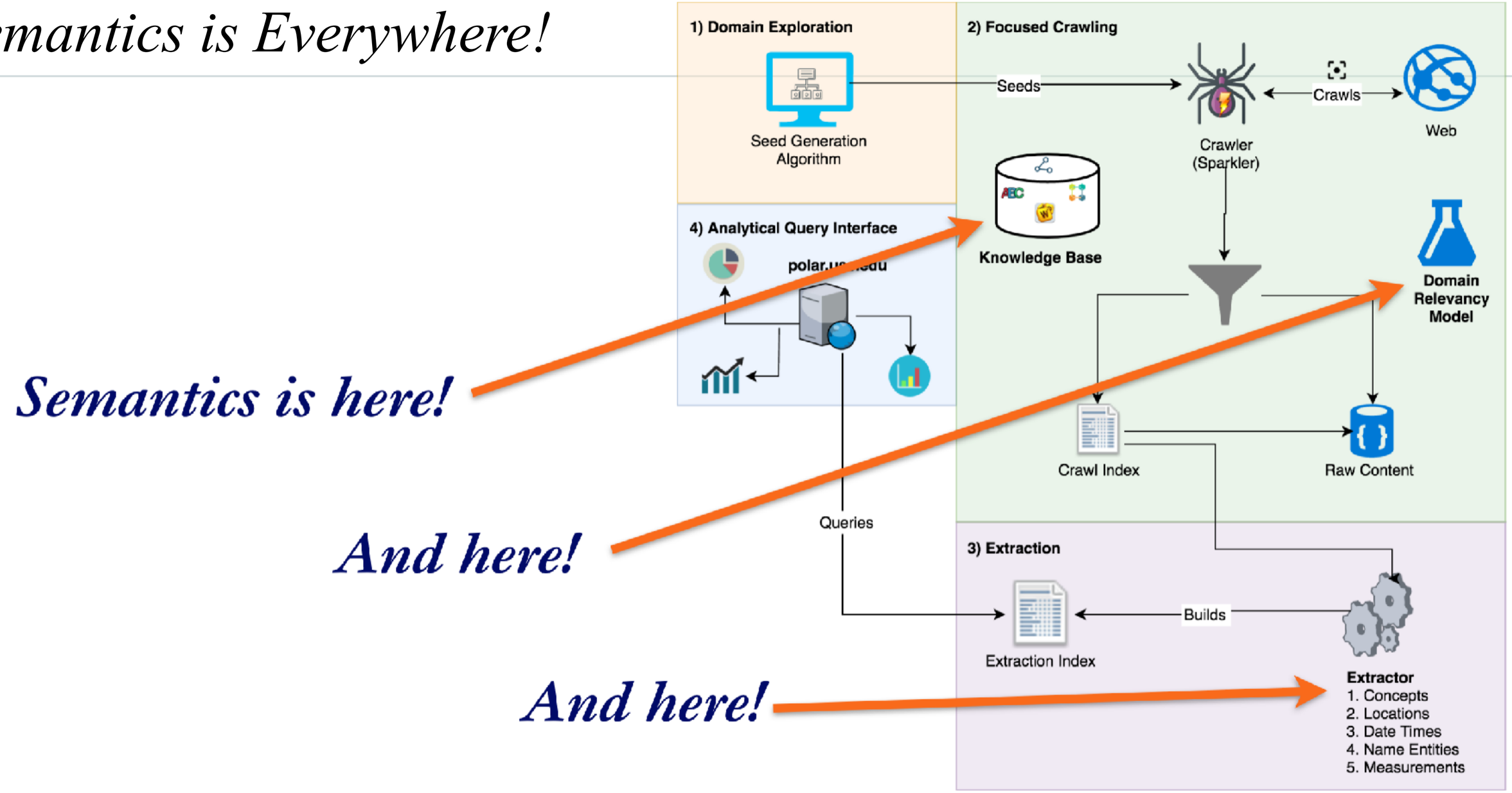
Query for documents mentioning the words “standards”, “ocean” and “sensors” using a keyword facet with terms “arctic” and “sea”

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”

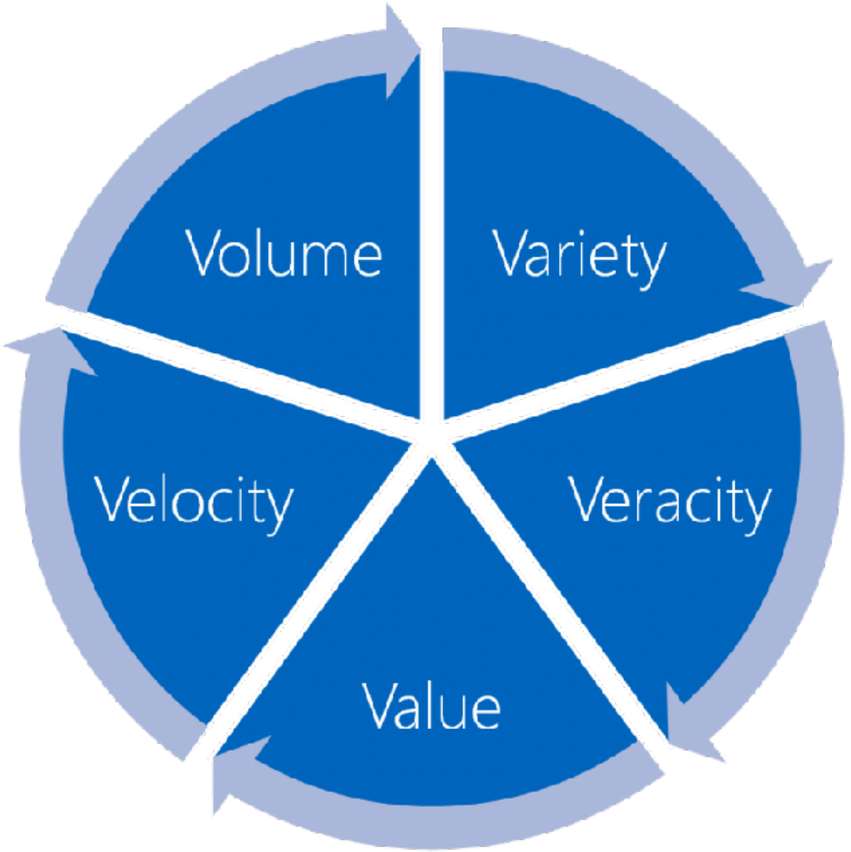


Semantics is Everywhere!



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