



Tailoring and Analyzing Topic-specific Crawls Using Semantics and Deep Learning

Ruth Duerr

Deep Learning and Semantics

Deep Learning for Data-Driven AI

Yoshua Bengio

Research Data Alliance 10th Plenary Meeting

19 September 2017



Université de Montréal



PLUG: Deep Learning. MIT Press book is out, chapters will remain online



Deep Learning for Data-Driven AI

Still Far from Human-Level AI

- Industrial successes mostly based on **supervised** learning



- Learning superficial clues, easy to fool trained networks
- Many more years of basic research needed



Deep Learning and Semantics

Deep Learning

Research

Still Far from Human-Level AI

- Industrial successes mostly based on **supervised** learning

data is the new oil

- Learning networks
- Many more
- Because AI is based on ML, successful AI applications require DATA – lots of data
- The first step in any project:
 - what data is available and what data is needed, do we need to collect more, do we need to label it?

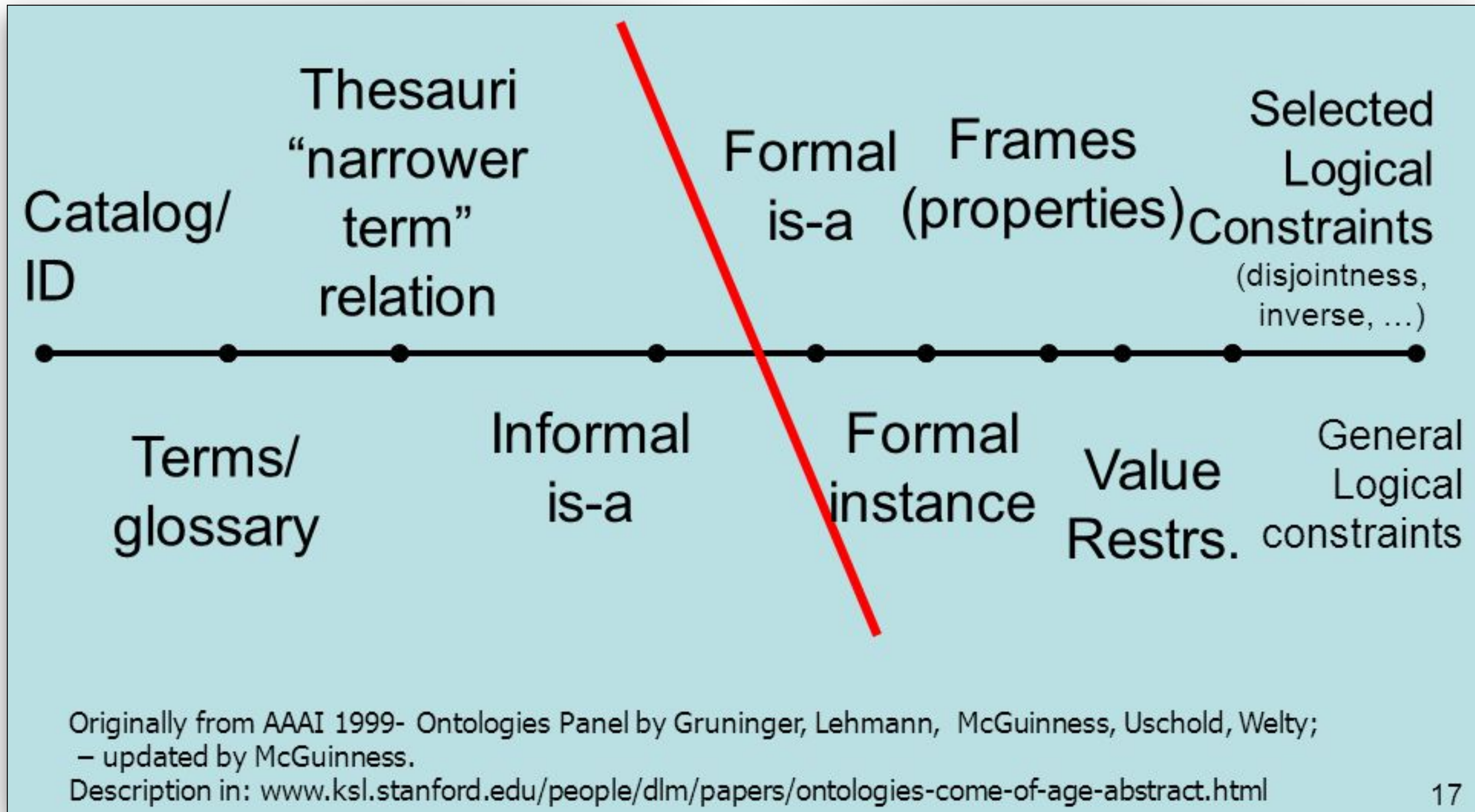
Université de Montréal

RD10

So, What is Labeled Data?

- There are infinitely many kinds
 - Nouns, verbs, adjectives, etc.
 - Entities (e.g., people, organizations, funders, etc.)
 - This image contains a hurricane (or person, or face, or animal, or building or whatever)

The Spectrum of Semantics - A Spectrum of Labeled Data?



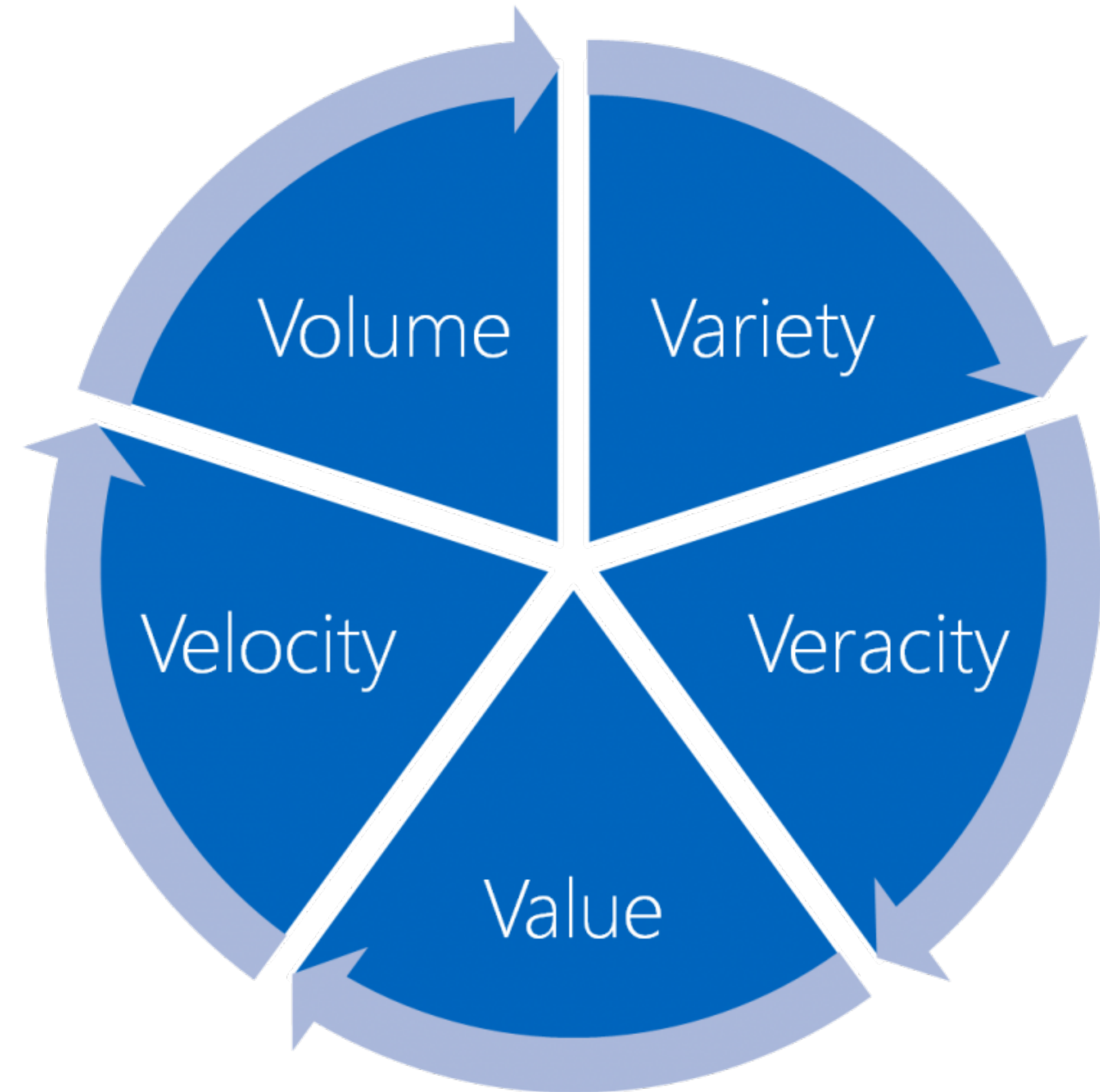
Originally from AAAI 1999- Ontologies Panel by Gruninger, Lehmann, McGuinness, Uschold, Welty;
– updated by McGuinness.

Description in: www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-abstract.html



So, What's the Problem?

- Domain 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*?



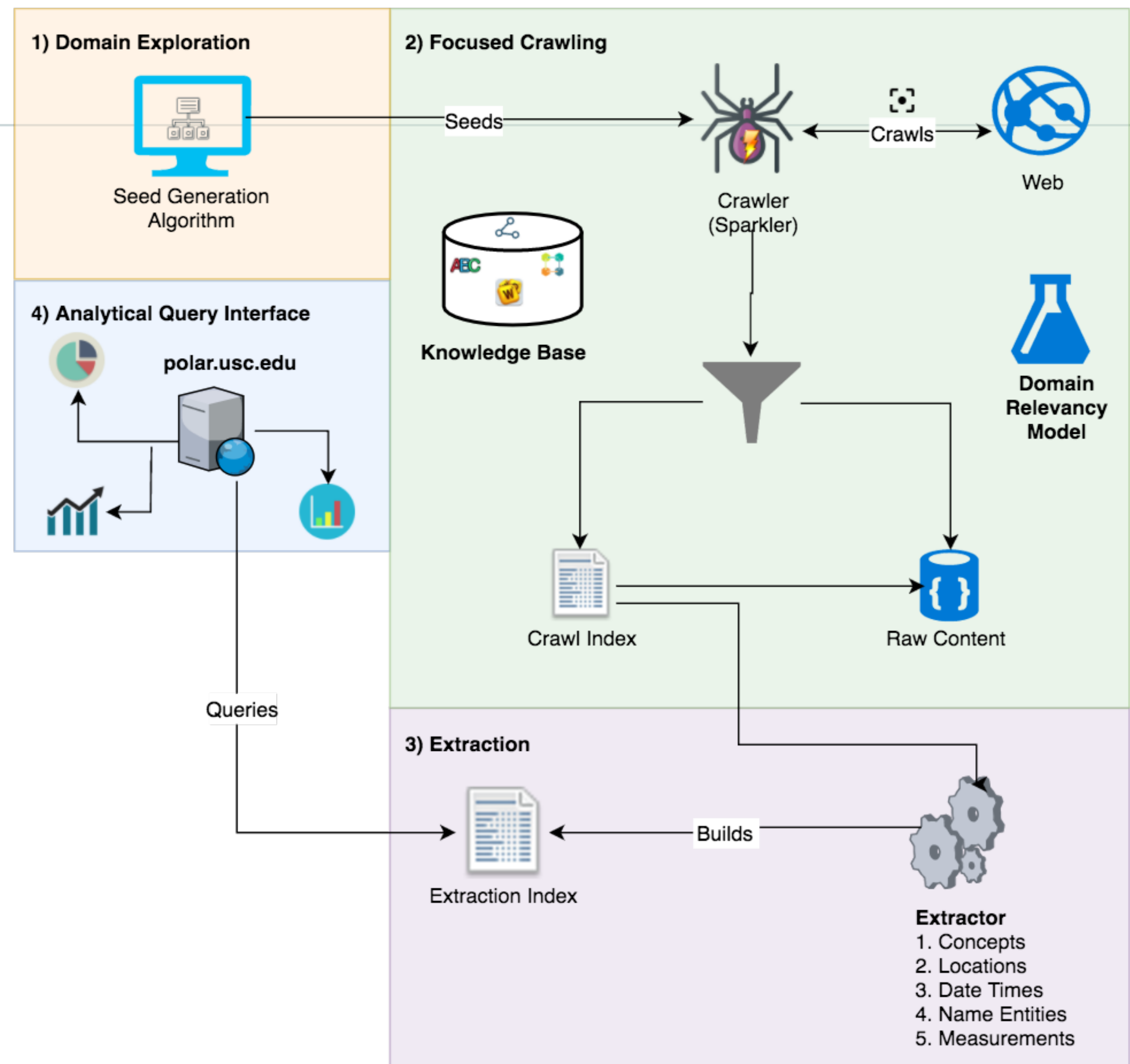
Applying “Big Data” Technology to Domain (here Polar) Data

- 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



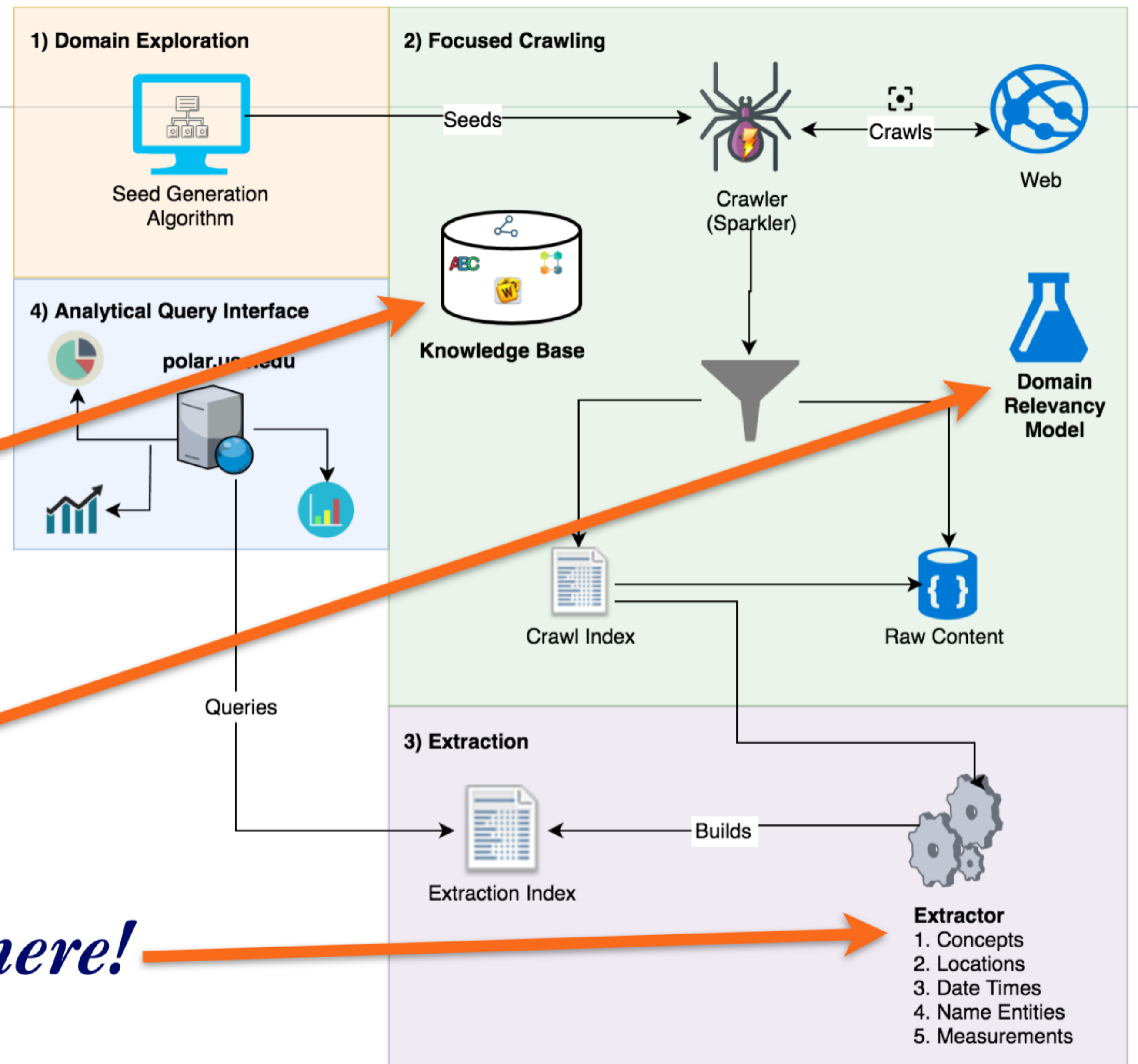
Domain Discovery System

Semantics is Everywhere!

Semantics is here!

And here!

And here!



Two Use Cases with Semantics in use with crawling and deep learning

- Sea ice use case - what insights can we get by simply applying domain expertise to a mostly existing set of tools
- Next steps with Ocean Observing Best Practices



Sea Ice Use Case - Crawler Inputs

Glossaries

http://pubs.usgs.gov/of/2004/1216/text.html	glacier terminology
http://unesdoc.unesco.org/images/0019/001925/192525E.pdf	glacier mass balance and related terms
http://www.aineva.it/previsori/Classificazione%20Internazionale/20080720_iacs_classif2008.pdf	snow terminology
https://nsidc.org/fgdc/glossary/	permafrost glossary (several pdf files required from website)
http://nsidc.org/cryosphere/glossary/all	general cryospheric glossary
http://www.jcomm.info/components/com_oe/oe.php?task=download&id=27226&version=March%202014&lang=1&format=1	sea ice glossary
http://globalcryospherewatch.org/reference/glossary.php	extensive compilation of cryospheric terms, most likely encompasses all of the above
http://www.spri.cam.ac.uk/resources/directory/organisations/	extensive list of polar organizations of various sorts

Sea Ice Use Case - Crawler Inputs

URL Seed List

- <http://arcticportal.org/>
- <http://ipa.arcticportal.org/>
- <http://nsidc.org/data/>
- <https://arcticdata.io/>
- <https://www.data.gov/climate/arctic-data/>
- <https://www.bas.ac.uk/data/>

Search Terms

- sea ice
- permafrost
- glacier
- ice sheet
- polar
- Arctic
- Antarctic
- snow
- sea ice thickness
- navigability
- calving
- ice berg
- snow water equivalent
- mass balance
- albedo

- After an initial crawl, a sample of the resulting documents were characterized as relevant, irrelevant, or possibly relevant and used to re-train the model

Sea Ice Use Case - Banana based query and analysis

Polar Deep Insights Concept Editor Query Interface Configure

Application configuration Complete

TREC-DD-PDF TREC-DD-SAMPLE NSIDC-CRAWL

Field	Value
Elastic search endpoint	http://polar.usc.edu/elasticsearch
Elastic search extraction index	polar-deep-insights-complete
Elastic search extraction doc-type	docs
Elastic search measurements index	polar-measurements
Elastic search measurements doc-type	raw-measurements
Entity Count JSON path	http://polar.usc.edu/html/polar-deep
Sweet ontology path	http://polar.usc.edu/html/polar-deep

[SAVE](#)

A concept ontology is required for this application to function. You can [create](#) your own ontology or download a predefined ontology from your elasticsearch index.

[DOWNLOAD](#)

The analytics interface requires additional precomputed information (ie) document counts per-entity.

[DOWNLOAD](#)

You can Curate extracted measurements [here](#).

[Github](#) . [Wiki](#) from IRDS.USC.EDU



Sea Ice Use Case - Banana based query and analysis

The screenshot shows the 'Polar Deep Insights' configuration interface. At the top, there are navigation tabs for 'Concept Editor', 'Query Interface', and 'Configure'. Below this, the 'Application configuration' section is marked as 'Complete'. There are three tabs: 'TREC-DD-PDF', 'TREC-DD-SAMPLE', and 'NSIDC-CRAWL'. A table lists configuration fields and values:

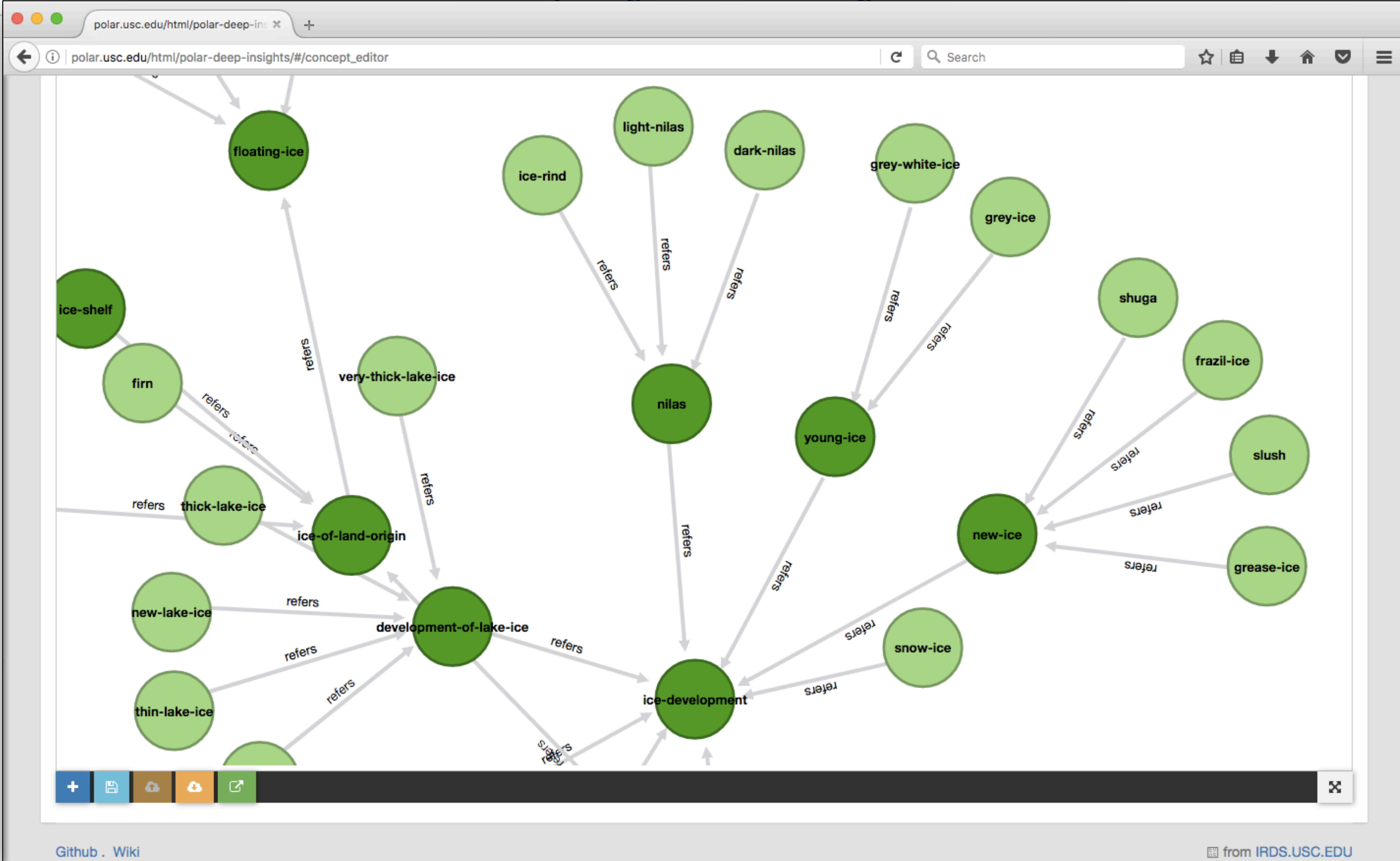
Field	Value
Elastic search endpoint	http://polar.usc.edu/elasticsearch
Elastic search extraction index	polar-deep-insights-complete
Sweet ontology path	http://polar.usc.edu/html/polar-deep

A red box highlights a message: 'A concept ontology is required for this application to function. You can create your own ontology or download a predefined ontology from your elasticsearch index.' Below this message is a 'DOWNLOAD' button. A red arrow points from this button to a larger, semi-transparent version of the same message and button overlaid on the configuration table.

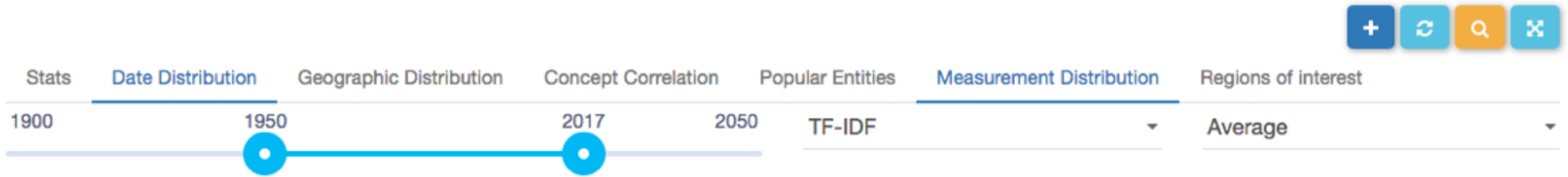
Footer: Github . Wiki | from IRDS.USC.EDU



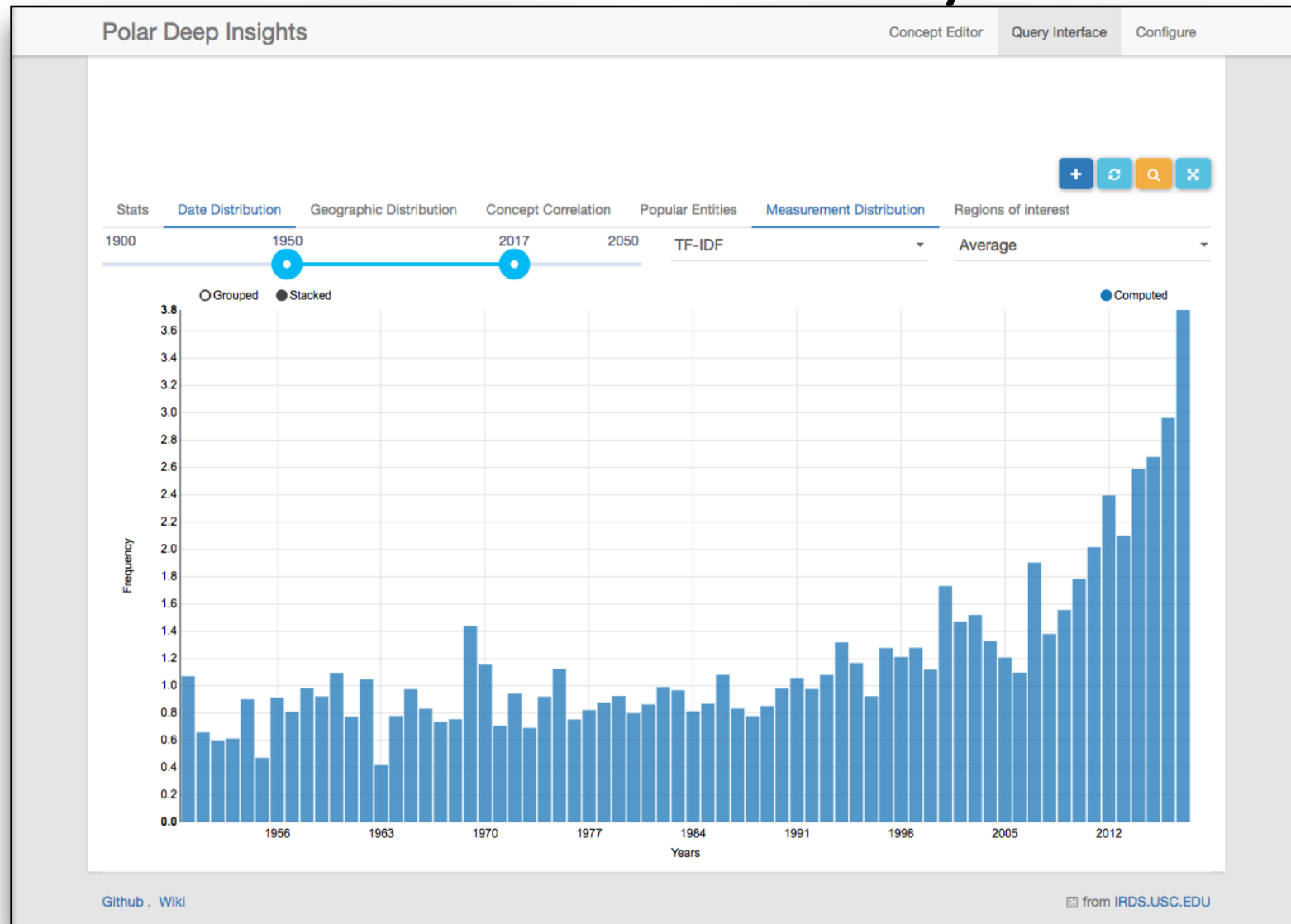
Sea Ice Use Case - Banana based query and analysis



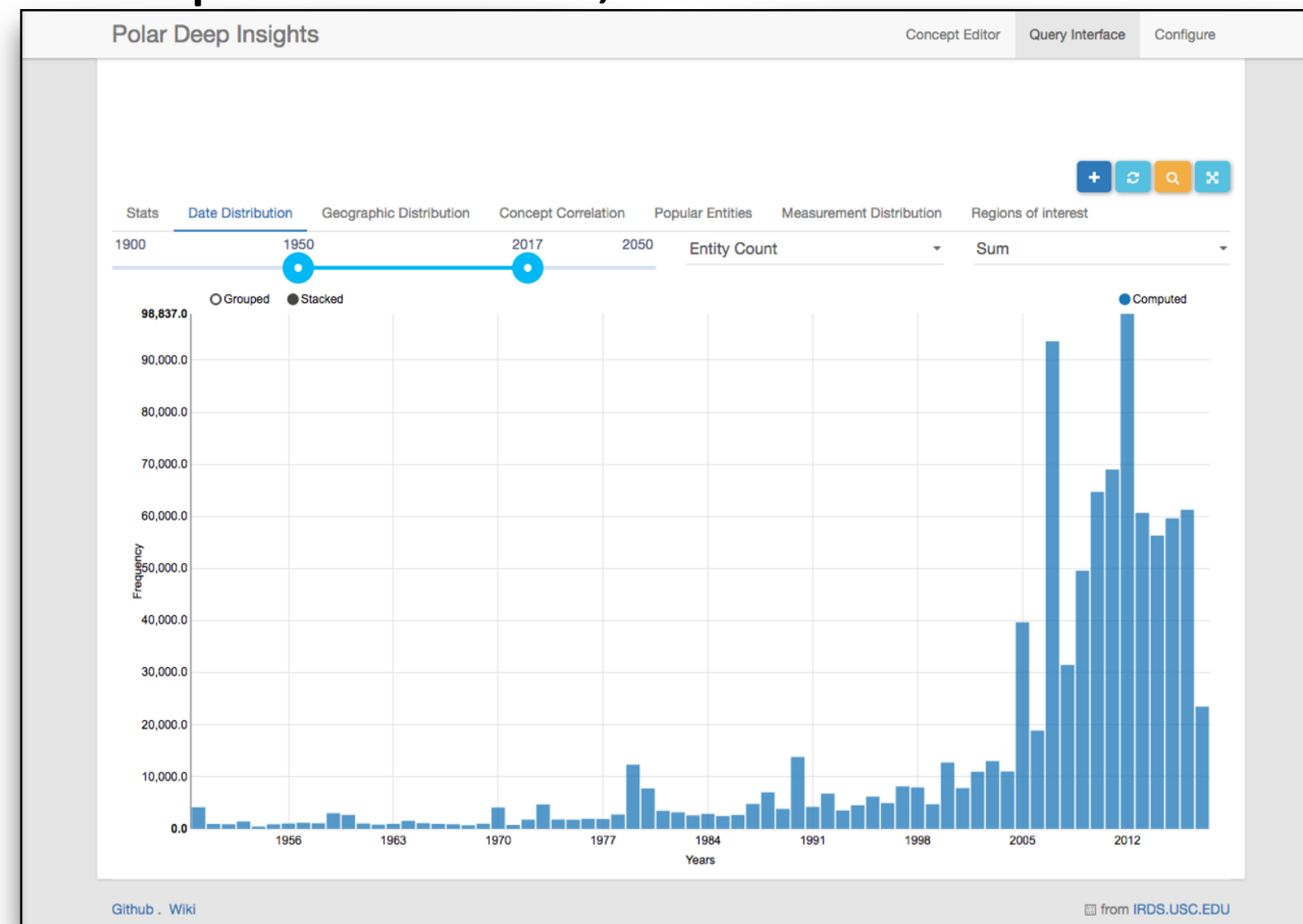
Sea Ice Use Case - Banana based query and analysis



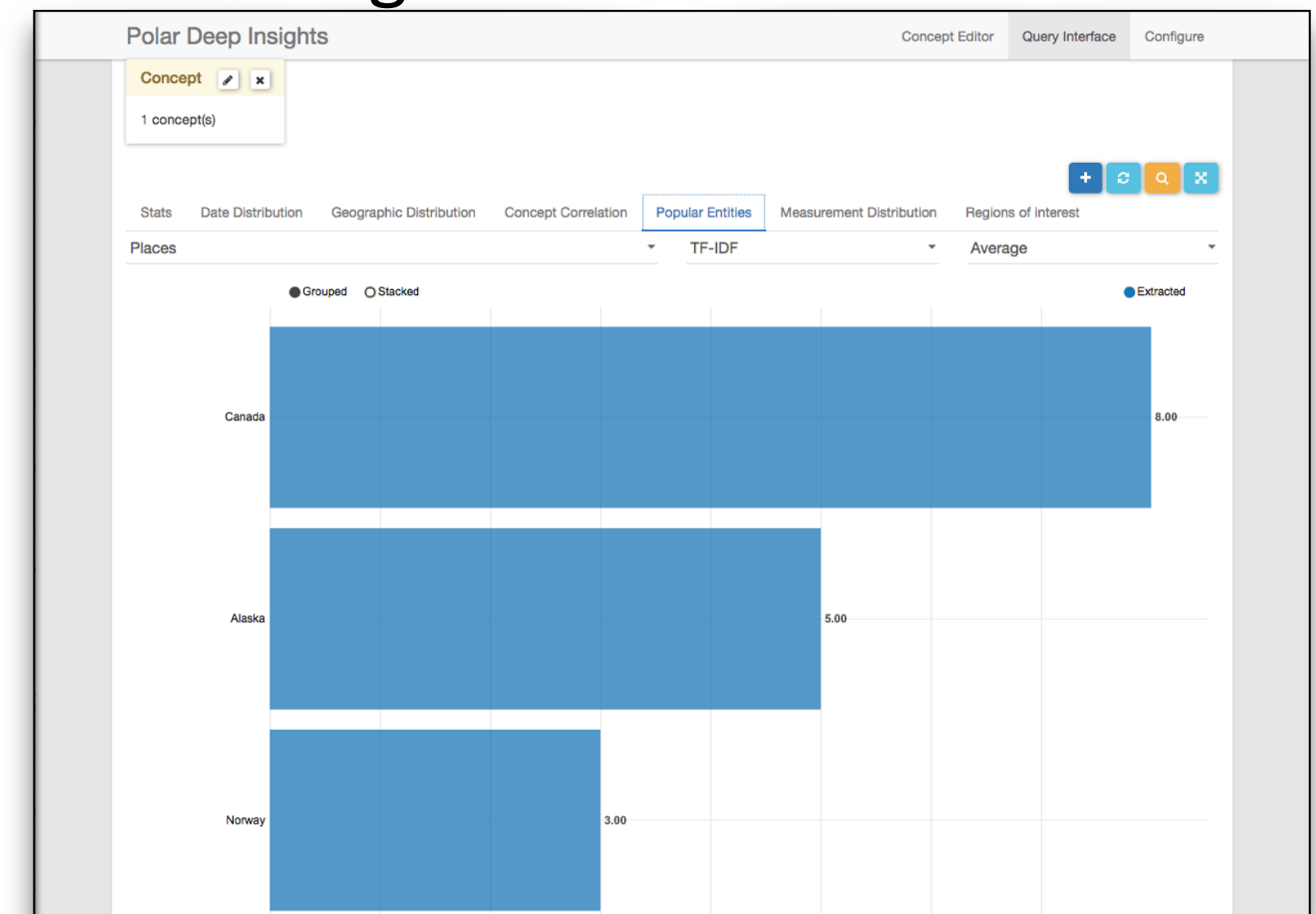
Distribution of documents by date



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.



Domain Exploration - Semi-automated Model Generation

Domain Discovery - Seed Generation Update Model

1 Generate a Model
ocean best practices
Minimum 10 each
78 66 108

More Options »

2 Create a Seed File

3 Start the Crawl

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/oce>

Guide to Best Practices for Ocean

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

Blocked by Content Security Policy

This page has a content security policy that prevents it from being loaded in this way.

Firefox prevented this page from loading in this way because the page has a content security policy that disallows it.

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

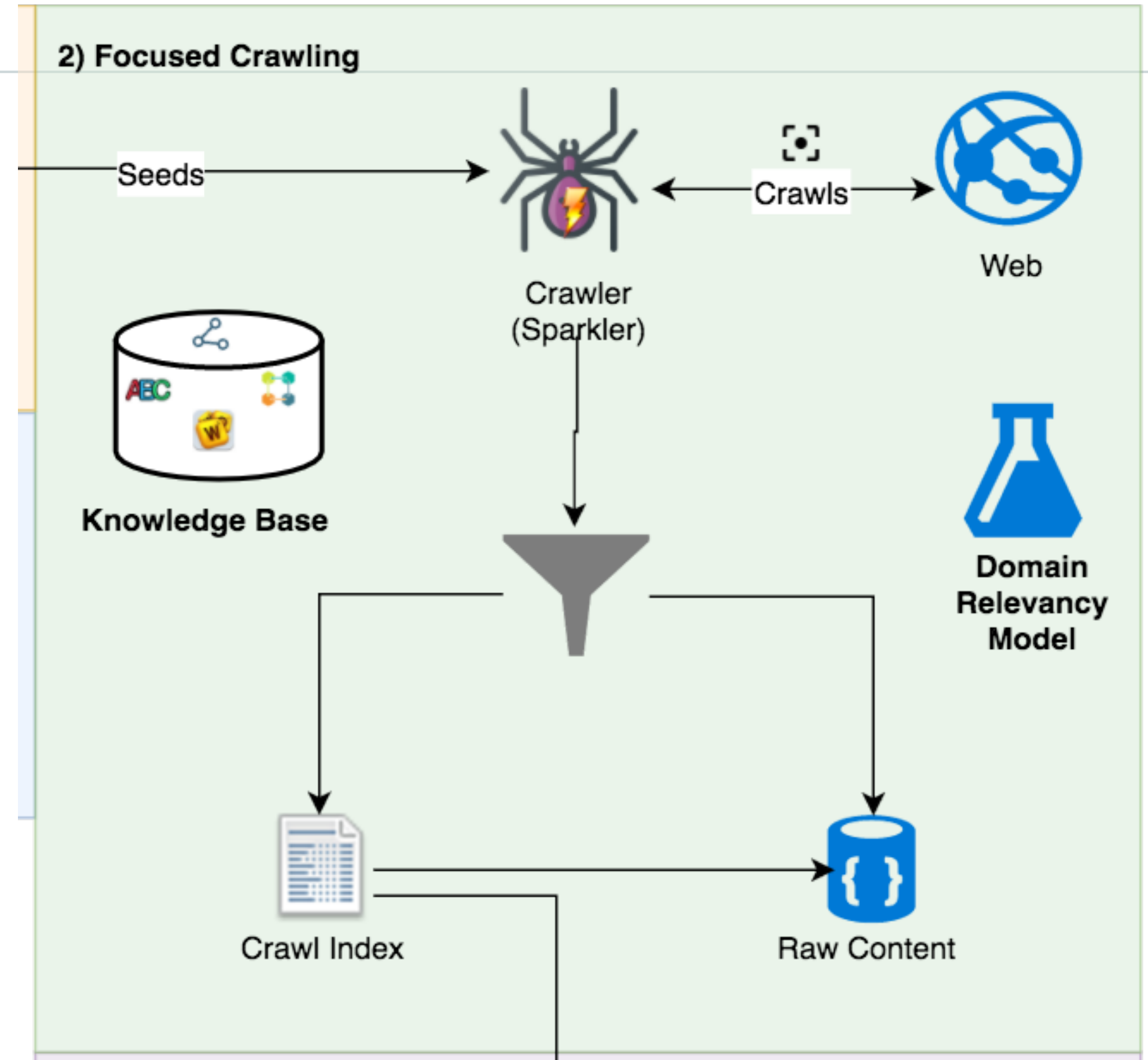
**The Ocean Data Standards and Best Pr
Project (ODSBP)**

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



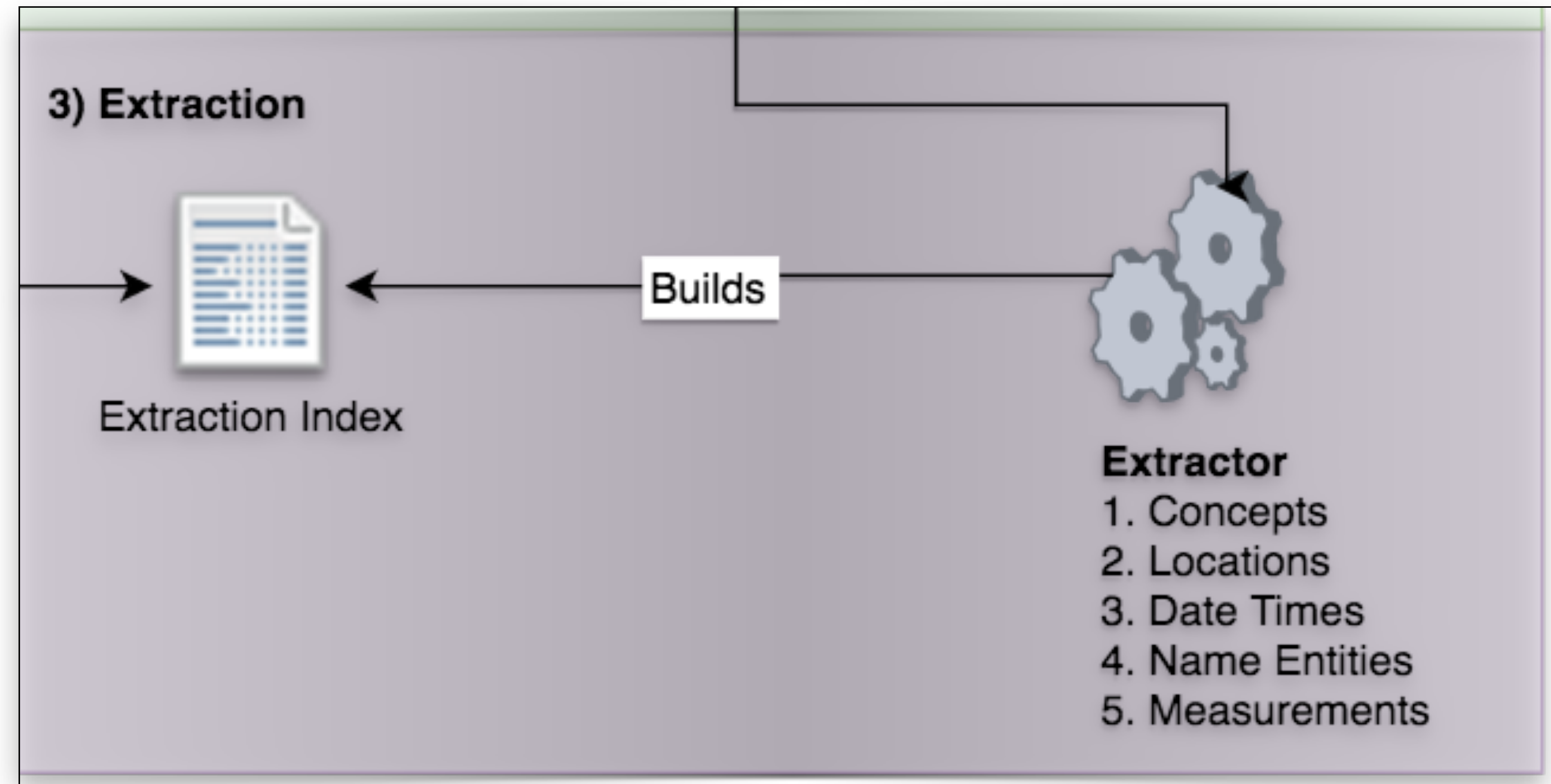
Focused Crawling

- 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



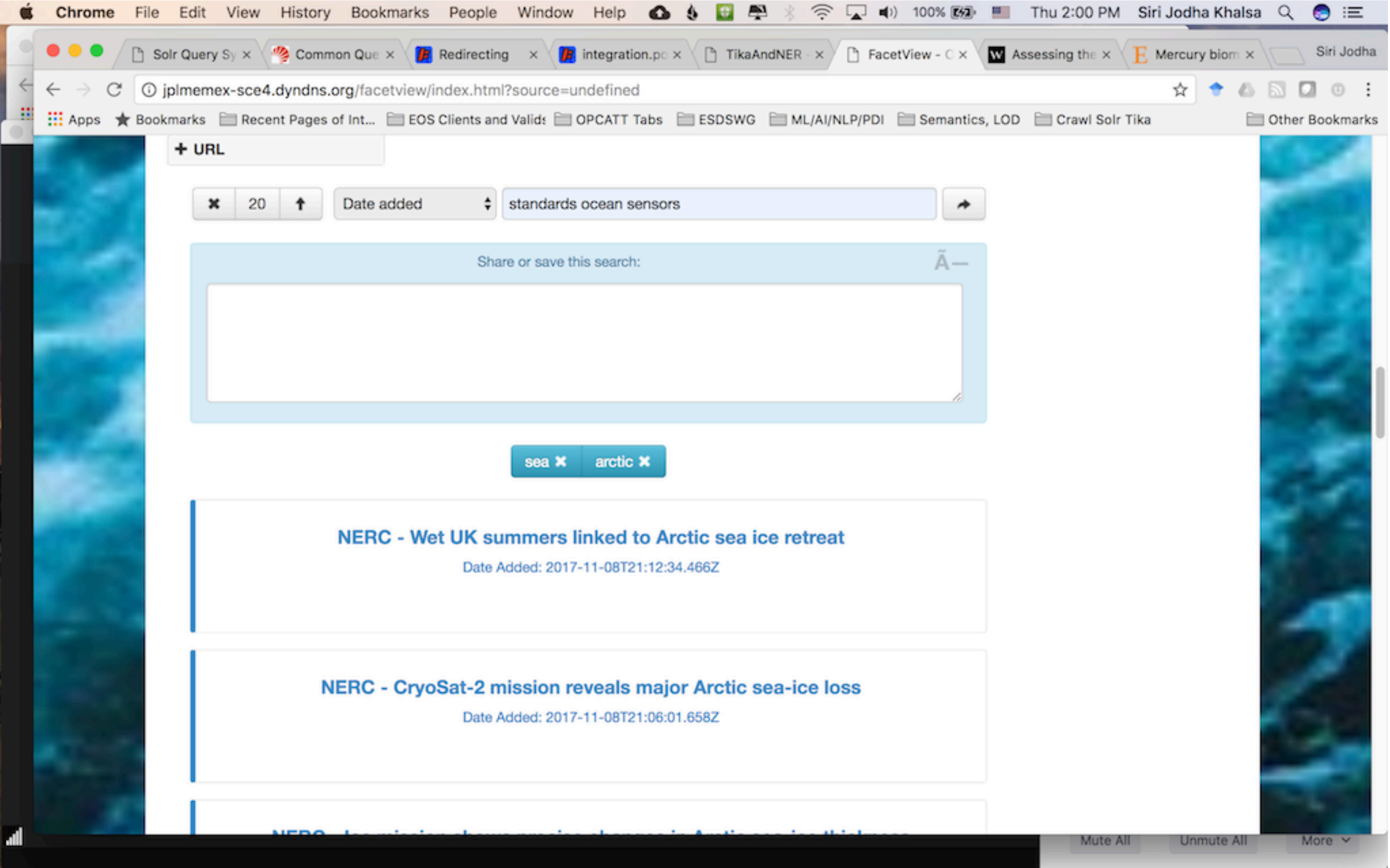
Extraction

- 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



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 document page with a table of contents and a search bar containing the word 'standards'. The rightmost window shows a document page with a search bar containing the word 'standards' and a table of contents. The document text includes sections on '2.5 Dissolved Organic Carbon Analysis' and '3 Results'.

2.5 Dissolved Organic Carbon Analysis

After the experiments, samples for DOC analyses were acidified (pH < 2) by addition of HCl and analyzed for nonpurgable organic carbon using a Shimadzu TOC-VCPH analyzer fitted with a Shimadzu ASI-V autosampler. 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. In addition to standards, aliquots of deep seawater reference material (Batch 10, Lot# 05-10) from the Consensus Reference Material (CRM) project were analyzed to ensure the precision and accuracy of the DOC analyses. Analyses of the CRM deviated by less than 5% from the reported value for these standards (41 to 44 μM DOC; <http://yyy.rsmas.miami.edu/groups/biogeochem/Table1.htm>). Standard and sample volumes analyzed were 20 to 40 mL. Routine minimum detection limits in the investigators laboratory using the above configuration are $2.8 \pm 0.3 \mu\text{M C}$, and standard errors are typically $1.7 \pm 0.5\%$ of the DOC concentration [Stubbins and Dittmar, 2012].

3 Results

3.1 Photodegradation of CDOM

Acknowledgements

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

