

A tool for constructing data albums for significant weather events

Data Albums
Compiled collections of information related to a specific science topic or event

Home About Contact Team

What is in a Data Album?
It contains links to relevant data files (granules) from different instruments, tools and services for visualization and analysis; information about the event contained in news reports, images or videos to supplement research analysis; and literature references.

Value to Researchers?
To gather relevant data and information for case studies and climatology analysis is tedious and time consuming, current NASA Earth Science data systems are not suited to assemble multi-instrument, multi-mission datasets around specific events. Data albums address this need by automatically gathering all the information and data needed.

Underlying Technology?
Data Albums are auto-generated by a reusable and scalable Knowledge Synthesis Engine (KSE). KSE uses domain knowledge encoded as ontologies that captures not only what information and data resources are relevant but also the semantic relationships between the various fragments of information and data.

Saffir-Simpson Hurricane Scale

1	64-83 mph
2	84-95 mph
3	96-110 mph
4	111-129 mph
5	130-156 mph

Additional Classifications

Tropical storm	39-73 mph
Tropical depression	39-73 mph

Events Overview Based on Data and Information Richness Year 2011

Diagram illustrating the Events Overview Based on Data and Information Richness for the year 2011. The diagram shows a large circle containing several smaller circles representing events, grouped by color and size. The events are: rina, gert, harvey, irene, philippe, jose, emily, maria, lee, nate, don, cindy, emily, arlene, and trent.

© Information Technology and Systems Center 2012, The University of Alabama in Huntsville, Huntsville, AL 35899, (256) 824-6868, info@itsc.uah.edu

Research Team

- PI - Rahul Ramachandran (UAH)
- Cols – Scott Braun (GSFC), Helen Conover (UAH), Michael Goodman (MSFC), Brian Wilson (JPL), Brad Zavodsky (MSFC)
- Development Team
 - Ajinkya Kulkarni (Tech Lead)
 - Rohan Bakare (GRA)
 - Sabin Baysal (GRA)
 - Roshan Sainju (Graduated)
 - Manil Maskey (Fixer)
- Testing/Validation/Documentation
 - Xiang Li
 - Shannon Flynn (Undergrad)

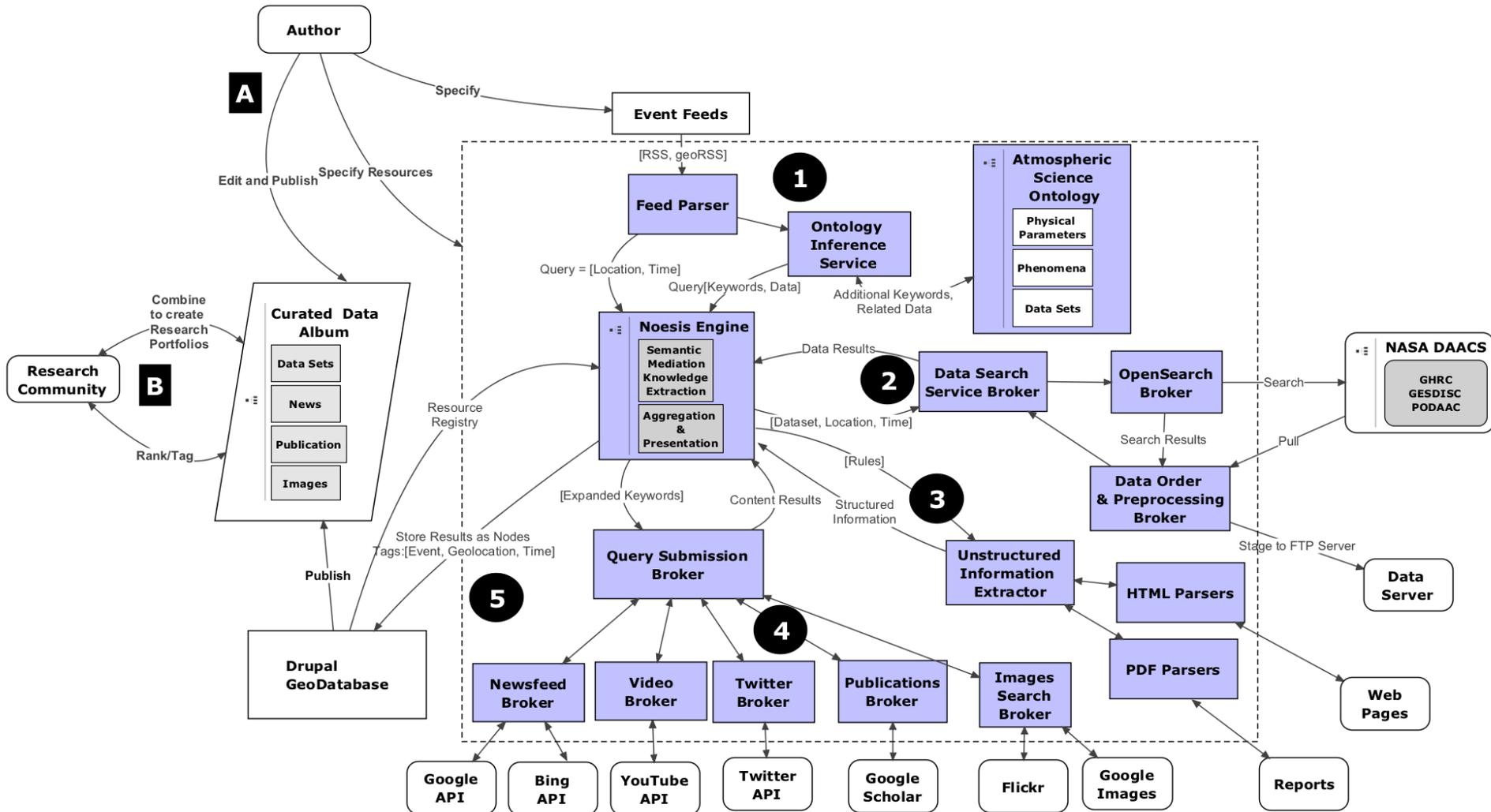
Curated Data Albums

- *Data Albums* are *compiled* collections of information related to a specific science topic or event with links to relevant data files (granules) from different instruments. Additional information include:
 - Tools and services for visualization and analysis
 - News reports, images or videos to supplement research analysis
- *Curation* provides the author of a Data Album the means to select the aggregated information

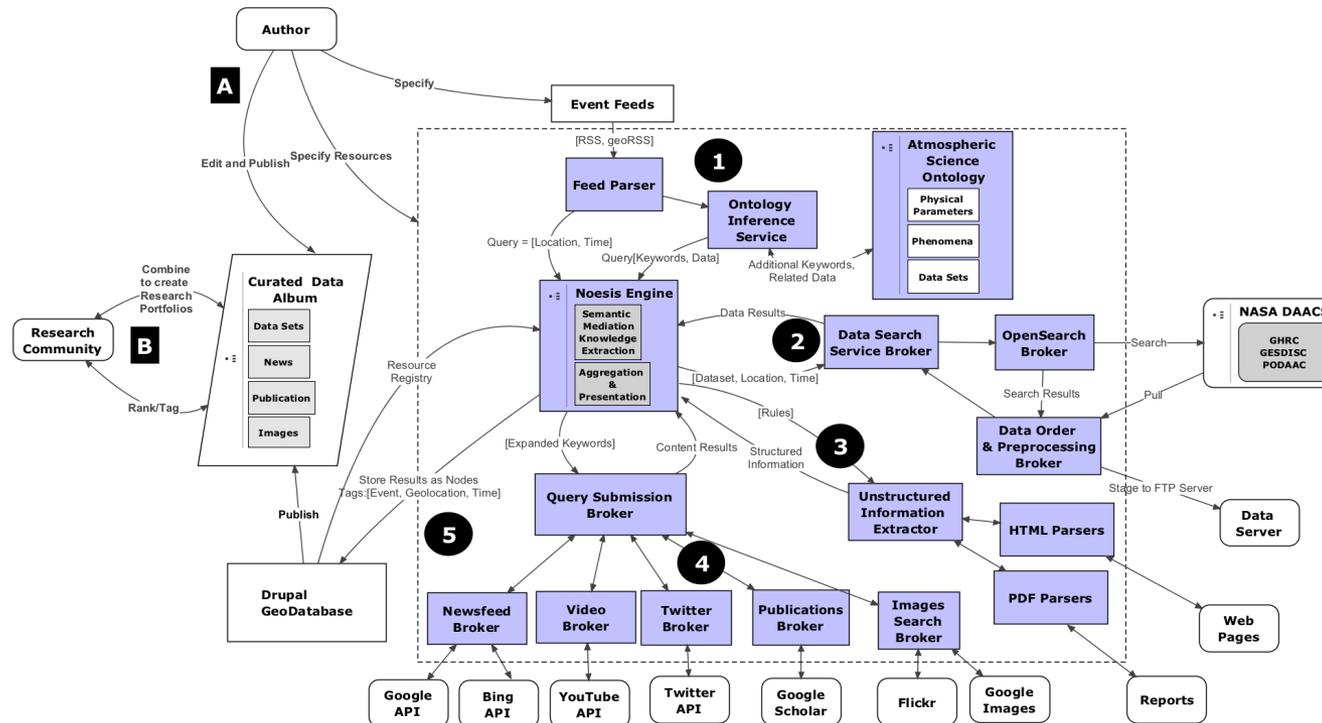
Motivation

- Case study analysis and climatology studies commonly used in Atmospheric Science research are instances where researchers studying a significant *event*:
 - Require data should be organized around the event rather than observing instruments
 - May need to discover new datasets that they would not have considered before
- Gathering relevant data and information for case studies and climatology analysis is ***tedious*** and ***time consuming***.
- Design of current Earth Science data systems assumes researchers access data primarily by instrument or geophysical parameter.
- *Need exists for tools to filter* through large volumes of online content and gather relevant information based on user's science needs.

Conceptual Information Architecture

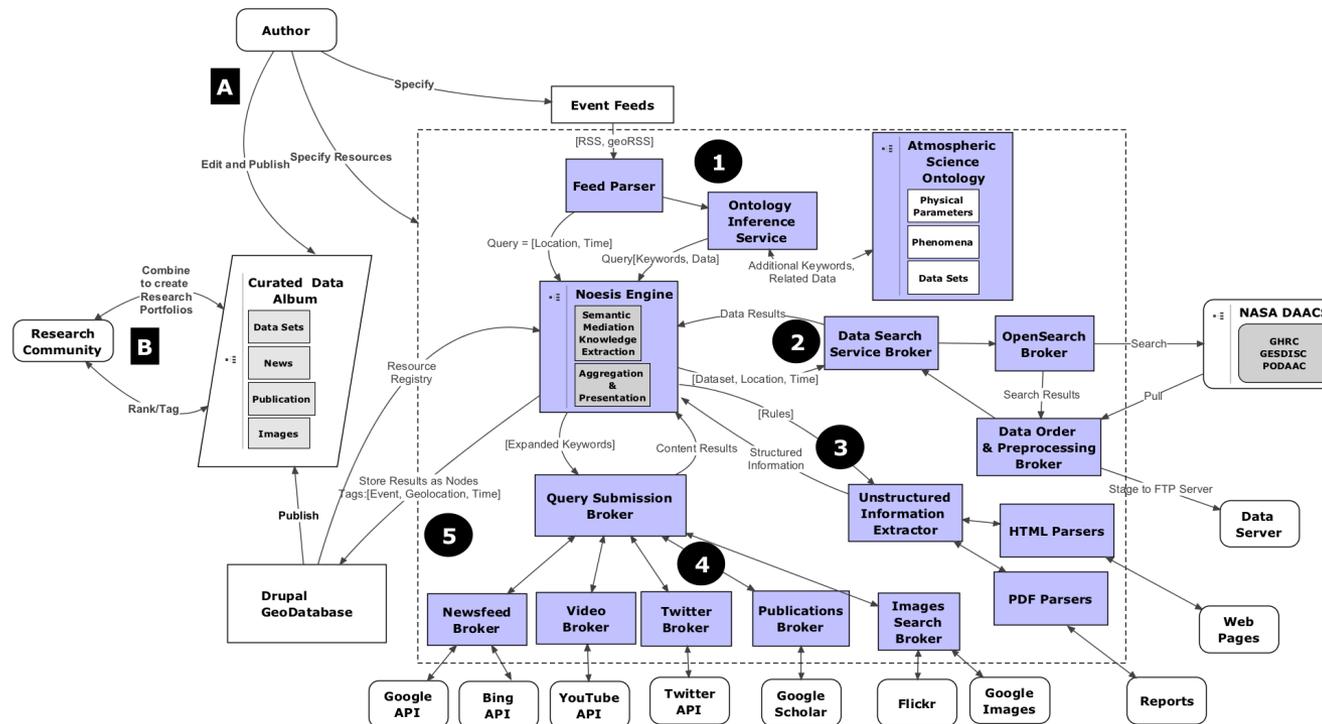


Component Interaction - 1



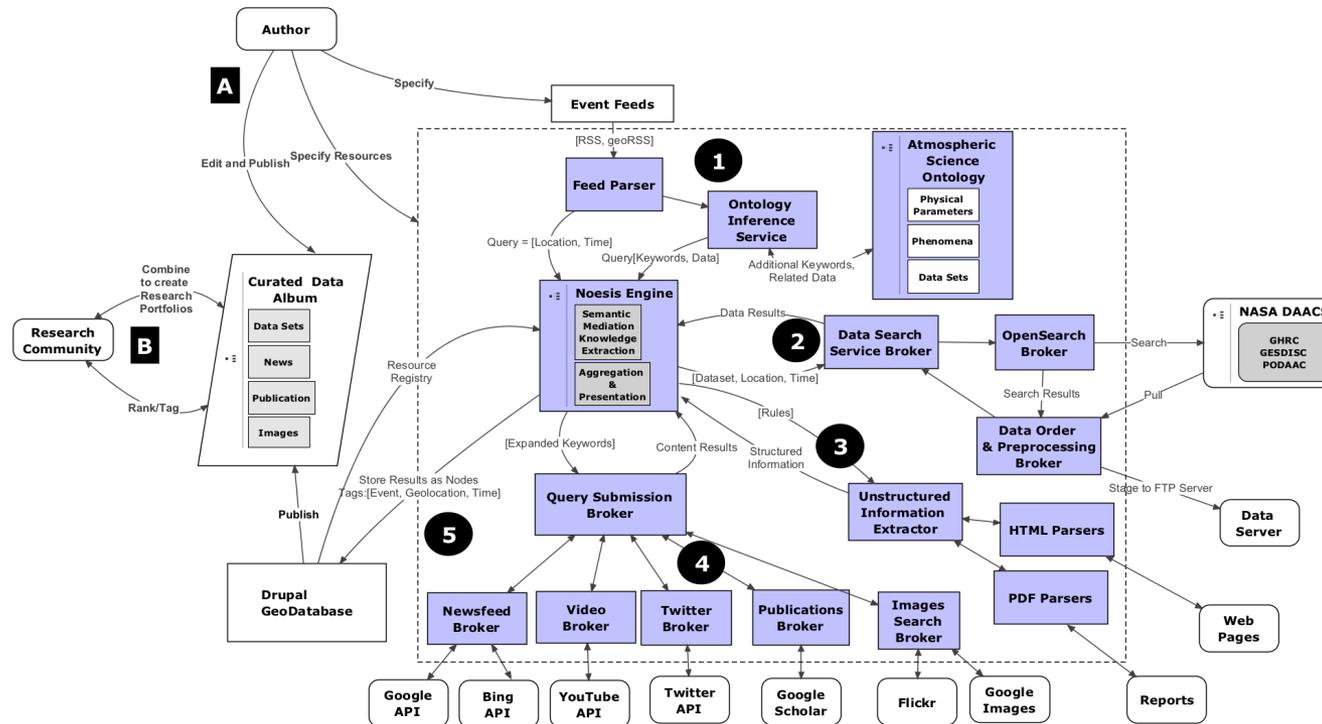
- Tool can be driven by an event feed. The feed parser will extract keywords, geolocation and time information. Keywords is fed to the OIS (Ontology Inference Service). The OIS provides additional contextual keywords including a list of relevant datasets to help refine the final query. These additional keywords, datasets, geolocation and temporal parameters are then fed to the Noesis Engine.

Component Interaction - 2



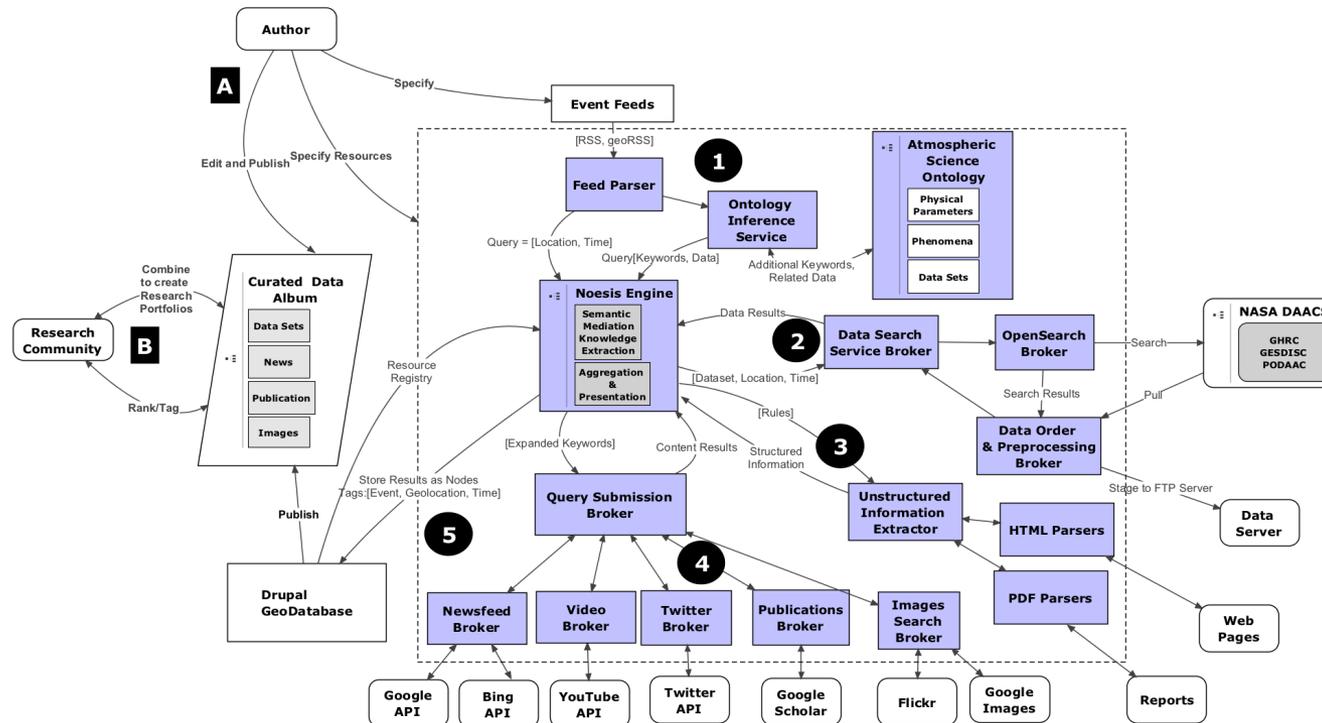
- Noesis Engine, responsible for framing the query for appropriate brokers, delegates the data search to the Data Search broker by providing dataset, keywords, location and time information. The Data Search broker then tasks an OpenSearch broker to obtain collection level metadata from the EOS Clearinghouse (ECHO) and NASA data centers, and return dataset descriptions.

Component Interaction - 3



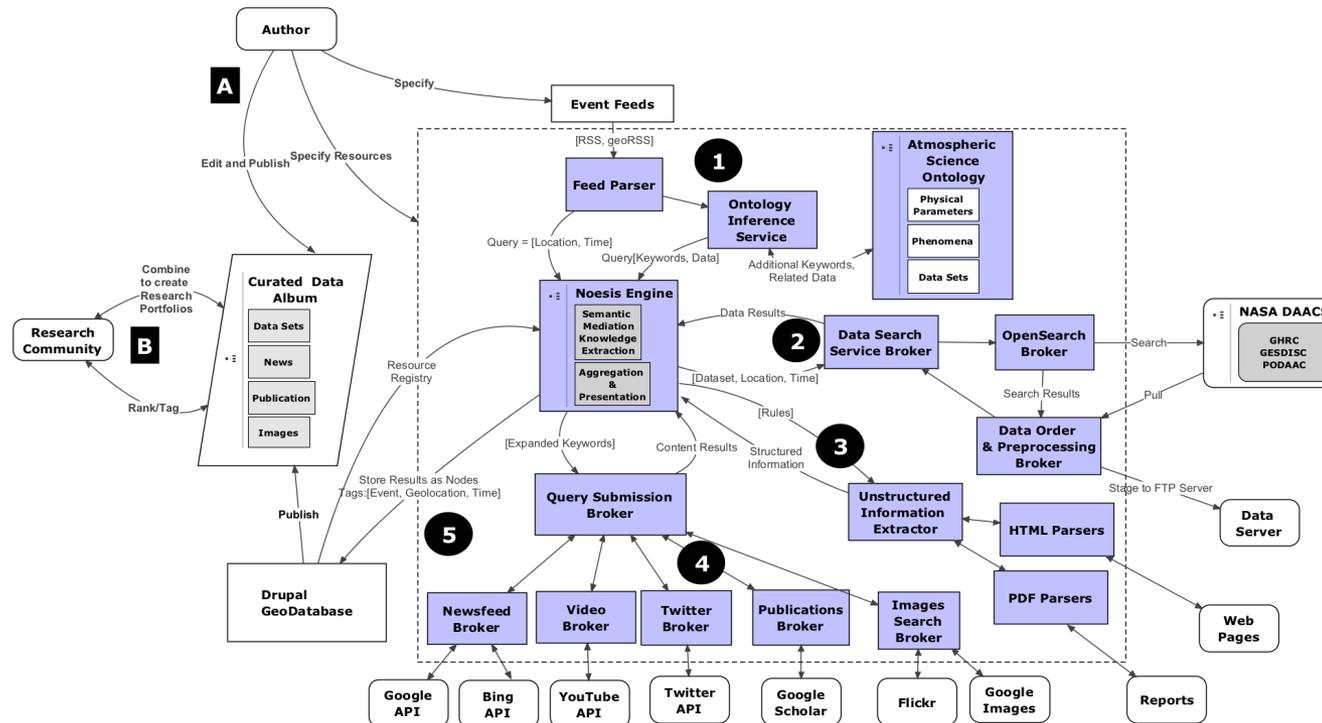
- Noesis Engine is also responsible for extracting information/knowledge from unstructured resources such as storm reports described in web pages and pdf documents. Noesis encodes “extraction rules” that are used to parse these documents to extract structured information and assemble these individual fragments into usable information

Component Interaction - 4



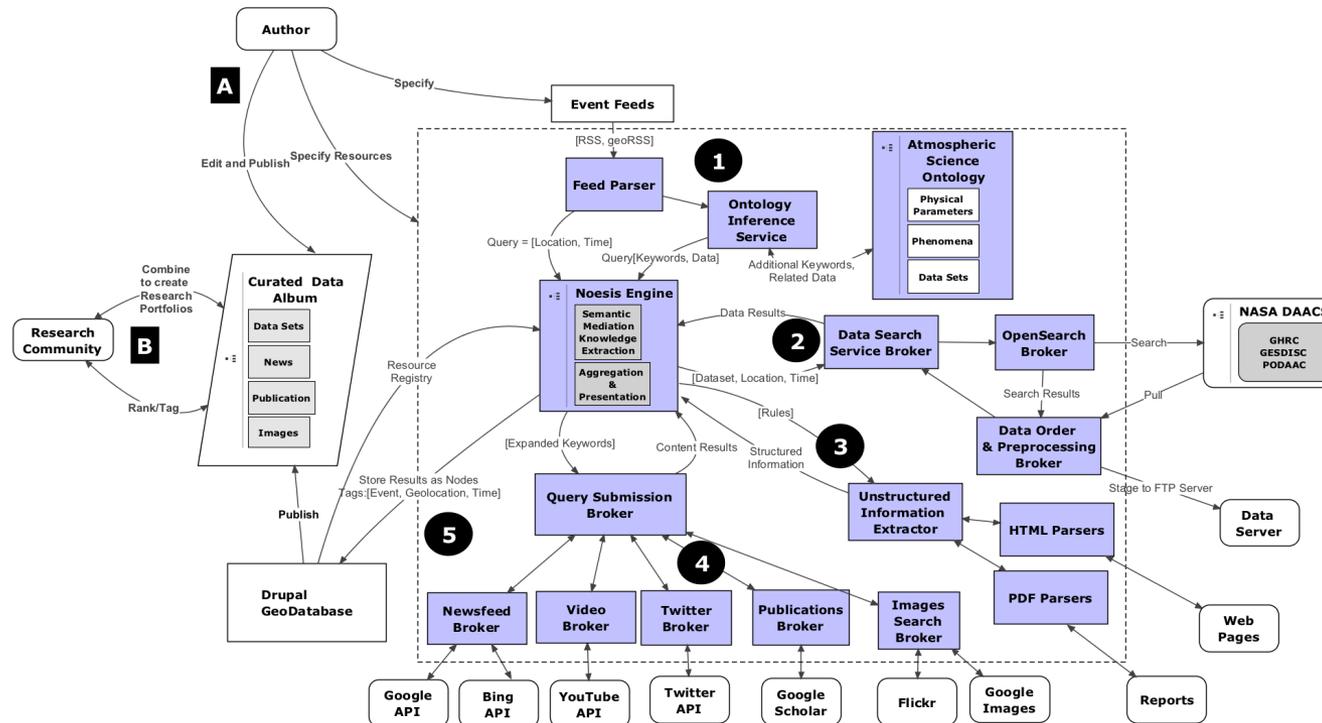
- Noesis Engine also feeds the expanded keywords to the Query Submission Broker to search for supplementary information potentially useful for research. The Query Submission Broker tasks individual brokers for different online content repositories, including Google API for news, YouTube for videos, Google scholar for any relevant published articles, Twitter and Flickr and Google Images for interesting photos.

Component Interaction - 5



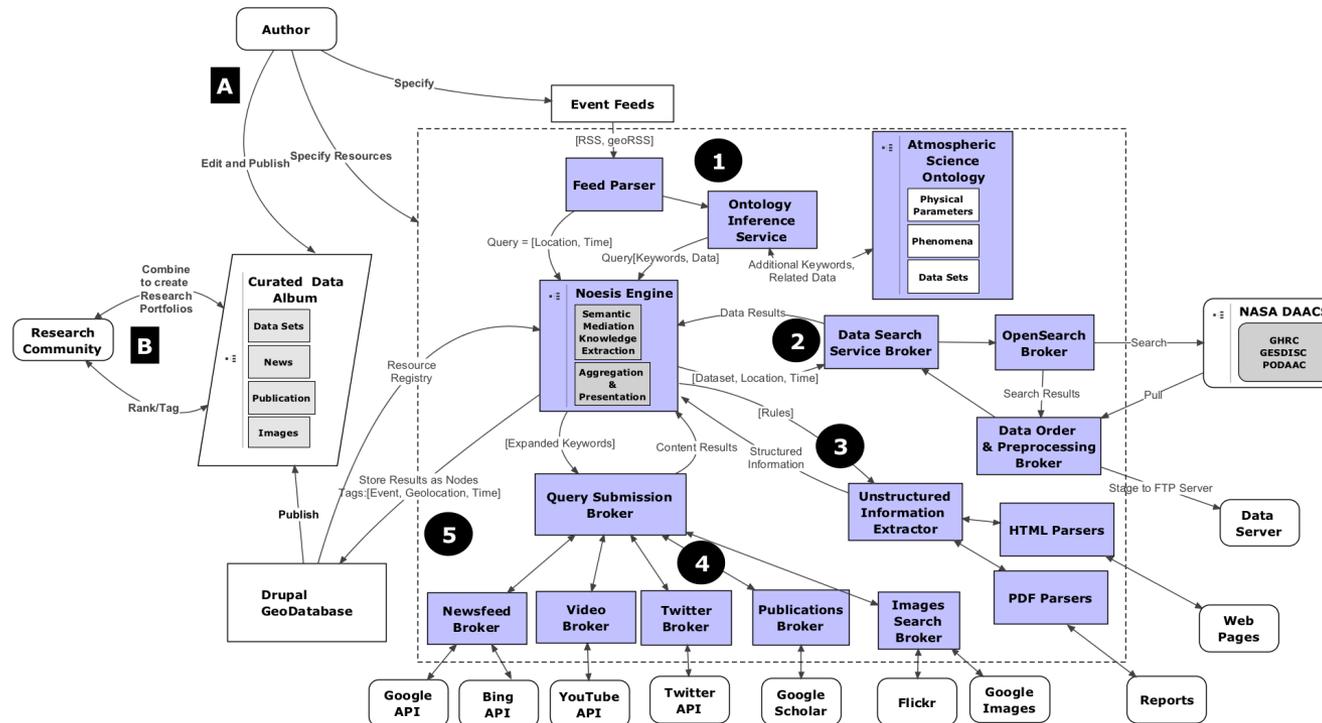
- Noesis Engine combines the aggregated data and content results and map this information into a Data Album. The Data Album is implemented as a custom Drupal node stored in the Drupal Geodatabase. The node is also annotated by all the keywords used to perform the query including expanded keywords, dataset names, geolocation and time.

Component Interaction - A



- Author with the appropriate permissions will be able to indicate a new event feed to trigger the creation of Data Albums. The author will be able to specify the different digital and data repositories to be searched, an important personalization feature to empower curation capabilities for their Data Album. The Data Album can also be shared with rest of the research community on the portal, and published as an information cast.

Component Interaction - B



- Once users start creating and sharing Data Albums, an online research community on hurricane research will evolve. The community can use the portal to discover useful Data Albums and to rate or annotate published Data Albums with new tags. The users will be able to search for specific published Data Albums and aggregate a series of Albums to create Research Portfolios for climatology studies.

Data Album Implementations

- Catalog of Hurricane Case Studies at GHRC. Noesis 2.0 will be integrated into the GHRC to create a portal to support hurricane case study research. (*Prototype Ready*)
- Case study generator at NASA's SPoRT Center. One goal of Noesis 2.0 will be installed at SPoRT to help automate the selection of weather events and other information needed for evaluating the SPoRT's mesoscale configuration of the Weather Research and Forecasting (WRF) model for convective cases (*To be implemented*)

Catalog of Hurricanes: Conceptual Architecture

Track (HRDAT)

Storm Summary (NCDC)

Storm Reports (NHC)

GIS Data (NHC)

Advisory Archive (NHC)

Mission Data (AOML)

NASA Data (ECHO)

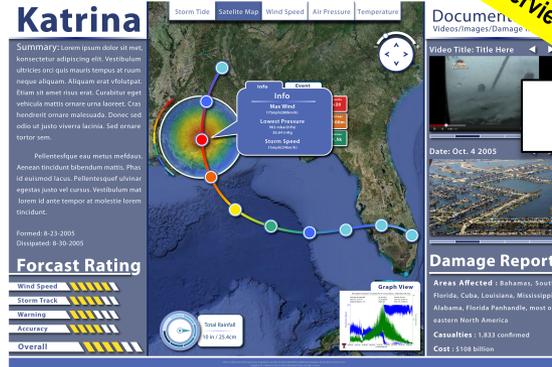
Pics/Videos (Wunderground)

Resources

NOESIS Knowledge Synthesis Engine

Structured Information

Tool



Auto-generated Overviews

Interactive Infographics

Curation, Customization

Storm Data Album

Search: []

My Storm Albums | My Preferences | New Videos | New Photos | New Storms

Storm Name: Katrina
Season: 2005

Album Details: About, Settings, Members

Data Albums

Search

Map View: Shows storm tracks over a geographical map with color-coded intensity levels.

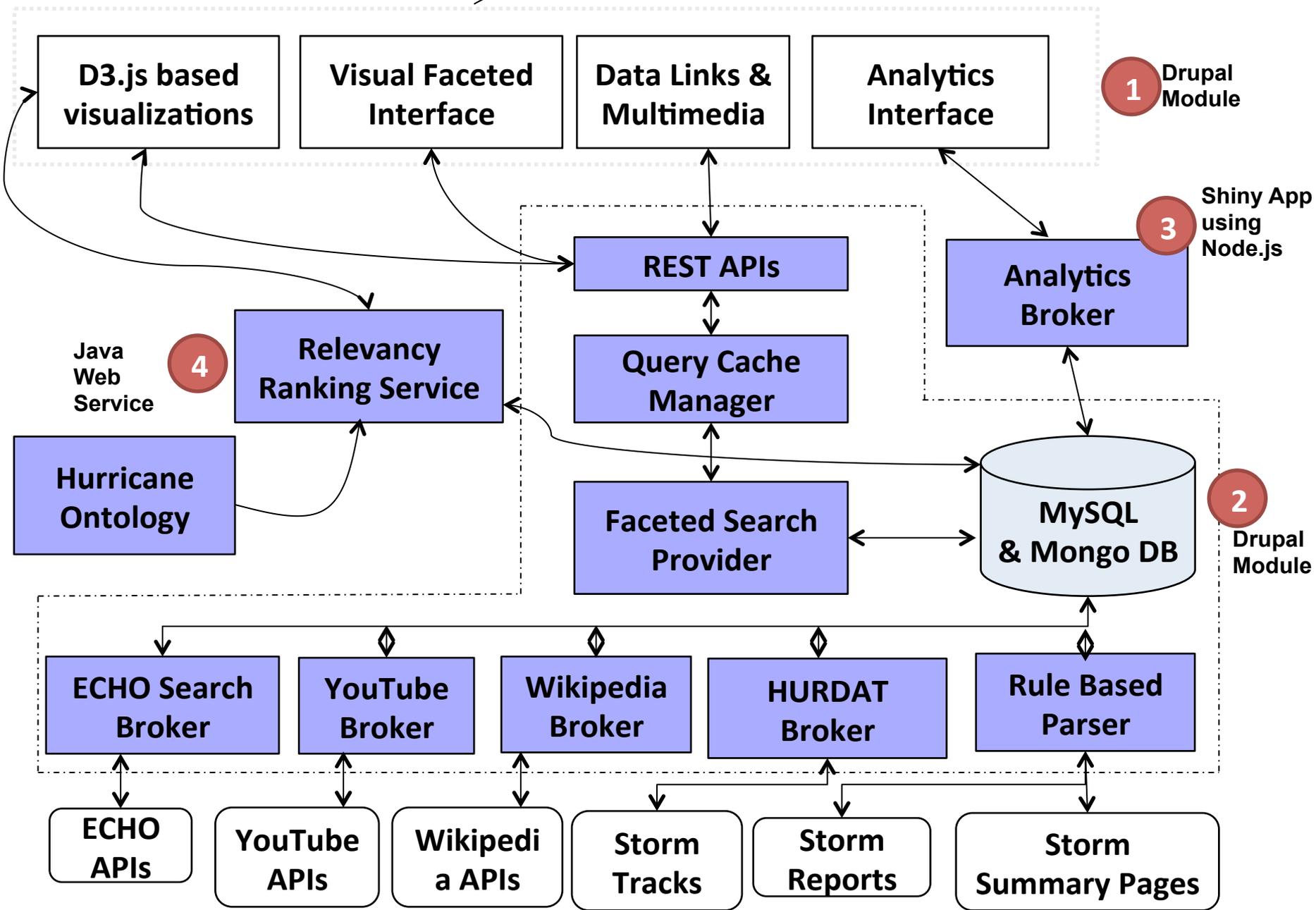
Search Results: List of search results including storm names, dates, and locations.

Discover, Explore

Views

Architecture Details

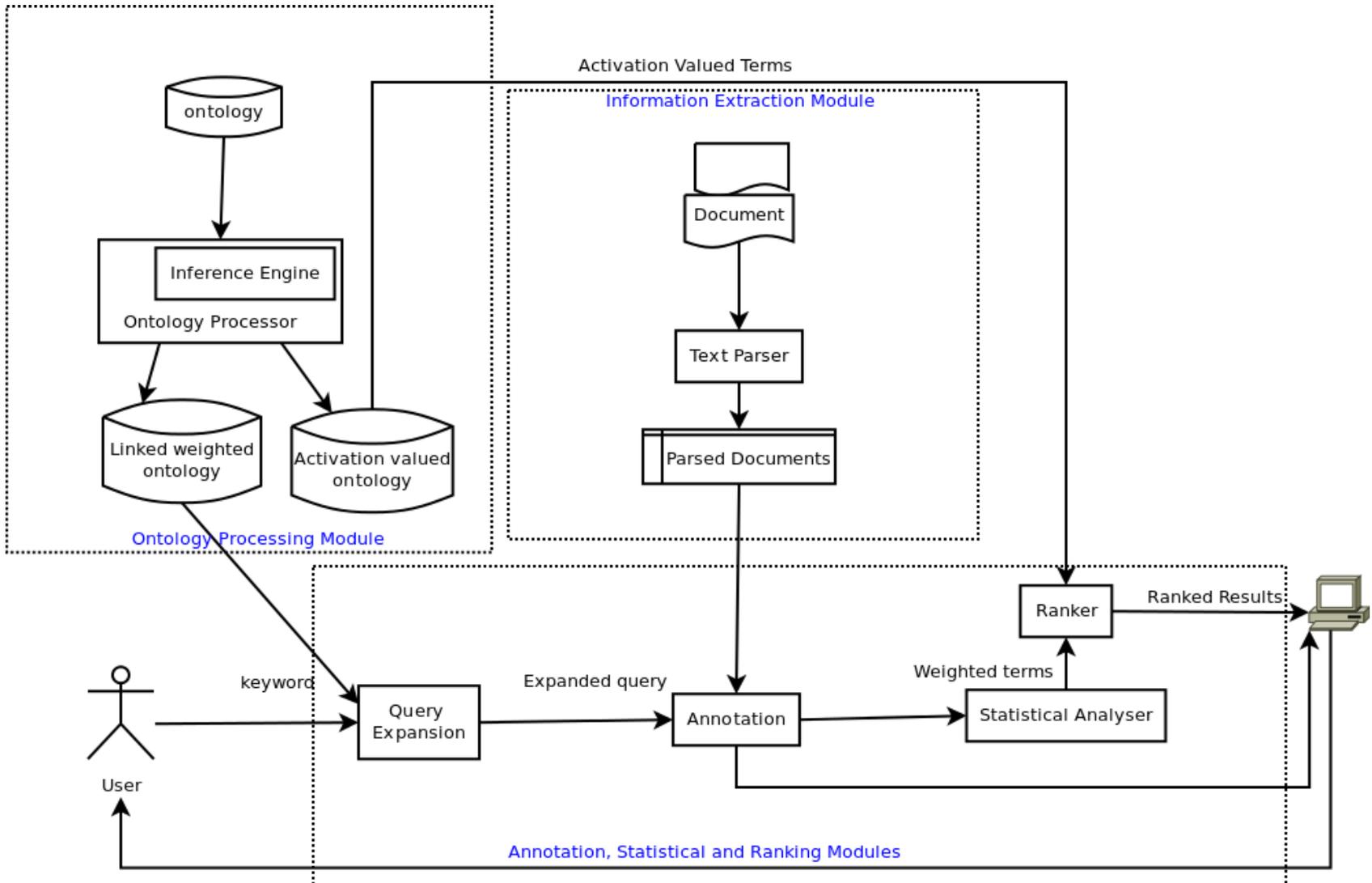
User



Ontology-based Relevancy Ranking Service

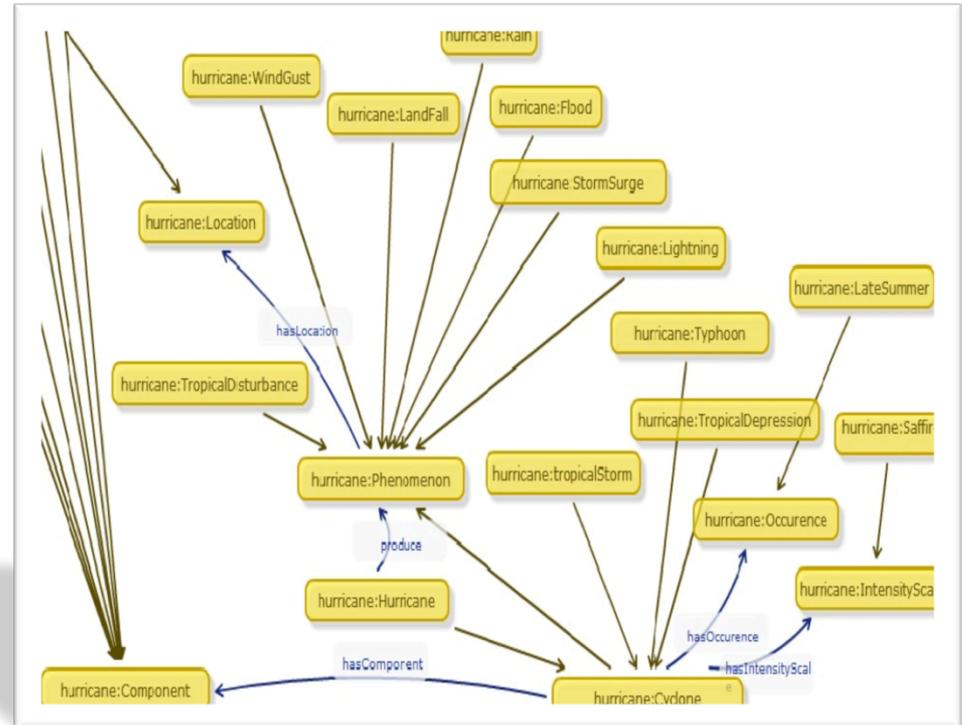
- General service that can be customized for specific applications
- Service needs to provide the following capability :
 - Feed application ontology to the system
 - Feed collections of documents that are need to be ranked
 - Provide search keywords
 - System provides the ranked documents.

Service Architecture



Initial Hurricane Ontology

- ▼ ● Ining
 - ▶ ● CasultyAndDamage
 - ▶ ● Component
 - ▶ ● Condition
 - ▶ ● DataSet
 - ▶ ● DataSourceOrigin
 - ▼ ● DataSourceType
 - ▶ ● Document
 - Image
 - Measurement = Observation
 - ModelSimulation
 - Observation = Measurement
 - Video
 - ▼ ● IntensityScale
 - Saffir-Simpson
 - ▼ ● Location
 - ▼ ● Tropics
 - Fall
 - TropicalIndianOcean
 - TropicalNorthAtlantic
 - TropicalWesternNorthPacific
 - ▼ ● Occurence
 - LateSummer
 - ▼ ● Parameter
 - AirPressure = Pressure
 - Pressure = AirPressure
 - Profile
 - RainRate
 - Speed = WindSpeed
 - StormSurgeHeight
 - ▼ ● Temperature
 - AirTemperature
 - SurfaceTemperature
 - Velocity
 - WindSpeed = Speed
 - ▼ ● Phenomenon
 - ▶ ● Cyclone
 - ▶ ● Flood
 - ▶ ● LandFall
 - ▶ ● Lightning
 - ▶ ● Rain
 - ▶ ● StormSurge
 - ▶ ● TropicalDisturbance
 - ▶ ● WindGust



Ontology-based Ranking Algorithm

- Algorithm combines ontology based and traditional statistical score to estimate relevancy of a resource
- Algorithm based on:
 - A. Bouramoul and M. Kholadi, “An ontology-based approach for semantic ranking of the web search engines results,” in *2012 International Conference on Multimedia Computing and Systems (ICMCS)*, 2012.
 - M. Shamsfard, A. Nematzadeh, and S. Motiee, “ORank: An Ontology Based System for Ranking Documents,” *International Journal of Computer Science*, vol. 1, no. 3, pp. 225–231, 2006.

Ranking Algorithm: Concept Weights

The weight of links between two concepts calculated with following equations

$$W(c_j, c_k) = \frac{\sum_{i=1}^m n_{i,j,k}}{\sum_{i=1}^o n_{i,j} + \sum_{i=1}^p n_{i,k}}$$

$W(C_j, C_k)$ is weight between C_j and C_k

$n(i,j,k)$ is total number of related concepts to both C_j and C_k

$n(i,j)$ is total number of related concepts to C_j

$n(i,k)$ is total number of related concepts to C_k

Weights are stored in database

Subject	Predicate	Object	wt
Hurricane	sibling	Typhoon	0.71
Hurricane	sibling	TropicalDepression	0.71
Hurricane	sibling	TropicalStorm	0.71
Cyclone	superClass	Phenomenon	0.44
Cyclone	subClass	Typhoon	0.33
Cyclone	subClass	Hurricane	0.33
Cyclone	subClass	TropicalDepression	0.33

Ranking Algorithm: Activation Value

- Spread Activation algorithm is used to calculate activation value for each concept in ontology.
- The concepts that are specified by user form the initial concept set. These initial concept sets are set to value 1.

$$I(c_j) = I(c_i) \times W(I(c_i), I(c_j))$$

- $I(c_j)$ is activation value of current concept
- $I(c_i)$ is activation value of previous concept
- $W(I(c_i), I(c_j))$ is link weight between two concepts
- Activation values are stored in database

Concept word (c)	Val (I)
Hurricane	1.0
TropicalDepression	0.71
TropicalStorm	0.71
Typhoon	0.71
Cyclone	0.33
Flood	0.19

Ranking Algorithm: Statistical Analysis

- Term weight: TF-IDF model is used to provide weight to each concept words present in the documents.
- TF (Term Frequency): It measures the frequency of each concept word (c) in each document (d)

$$tf(c, d) = \frac{f(c, d)}{|d|}$$

Where:

- $f(c, d)$ is frequency of concept word (c) in document (d)
- $|d|$ is count of words in document (d)

Ranking Algorithm: Statistical Analysis

- IDF (Inverse Document Frequency): measures the importance of concept word (c) in entire document collection (D).

$$idf(c, D) = \log\left(\frac{|D|}{df_c}\right)$$

- |D| is count of the corpus (D)
- df_c is document frequency of (c): the number of document that contain (c)
- The tf-idf for the concept word calculated as below:
$$tf.idf(c, d, D) = tf(c, d) \times idf(c, D)$$
 - $tf.idf(c, d, D)$ is the tf.idf for (c) in (d) that is in corpus (D)
- The obtained tf.idf saved in memory to use for the ranking calculation.

Ranking Algorithm: Relevancy Score

- Score (S_d) for document (d) in collection of documents (D) is as below:

$$S_d = \sum_{i=1}^m I(c_i) \times tf.idf(c_i, d, D)$$

- S_d is final score of document (d)
 - $\sum_{i=1}^m I(c_i) \times tf.idf(c_i, d, D)$ is sum of score of each matched concept word and is obtained by multiplying ontology-based score and the statistical score
- The obtained score is used to rank the document (D)

Algorithm Evaluation

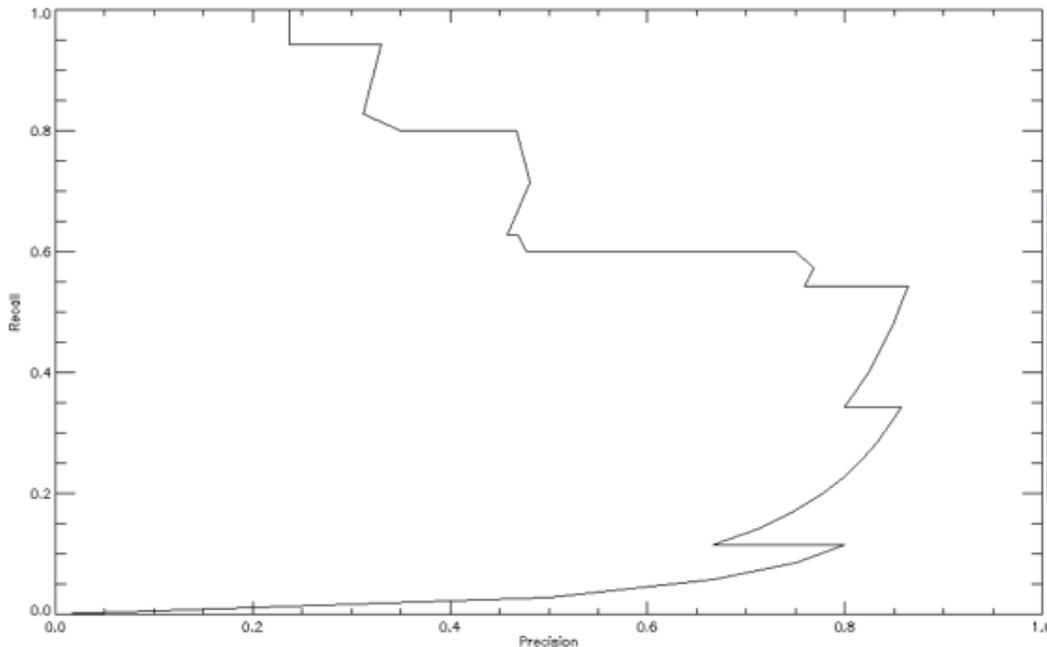
- Compared algorithm results versus “truth data”
- Truth data:
 - A total of 35 GHRC collections were manually selected
 - Considered as useful and relevant to hurricane studies
- Top 35 ranked results from the algorithm are used for performance evaluation.

	Yes (Algorithm)	No (Algorithm)
Yes (Manual)	21	14
No (Manual)	14	111

- $Accuracy = (21+111)/160 = 82.5\%$
- $Precision = 21/35 = 0.60$
- $Recall = 21/35 = 0.60$

Algorithm Application

- Precision vs Recall Graph



Relevancy Score Threshold : 0.0140
Identified number of collections: 60.0
Matched number of collection: 28.0
Precision: 0.466667
Recall: 0.800000

- Ideally you want high precision, high recall
- For data search, we want higher recall:
 - Gets everything that is important
 - Gets non-relevant data (users can filter out)

Algorithm Improvements

- Improving the ontology richness
- Utilizing semantic relationships to improve weight calculations
- Utilizing concept word “location” to improve weight calculations.
 - Example: Keywords vs Description

Key Features

- Autogenerates infographic view of Hurricanes
- Visual faceted search interface for Hurricanes
 - Results ordered alphabetically
 - Storms color coded by category
- Visual faceted search interface data for a given Hurricane
 - Data collection color coded by relevancy
 - Relevancy thresholds can be adjusted by the user
- Analytics
 - Summary statistics for events such as box plots, density map
 - Interactive plots that allow user to configure and display six parameters within a scatter plot
- Download scripts to facilitate access to the data files

Version 2 Snapshots

Data Albums

Compiled collections of information related to a specific science topic or event

About the Data Albums

What is in a Data Album?

It contains links to relevant data files (granules) from different instruments, tools and services for visualization and analysis; information about the event contained in news reports, images or videos to supplement research analysis, and literature references.

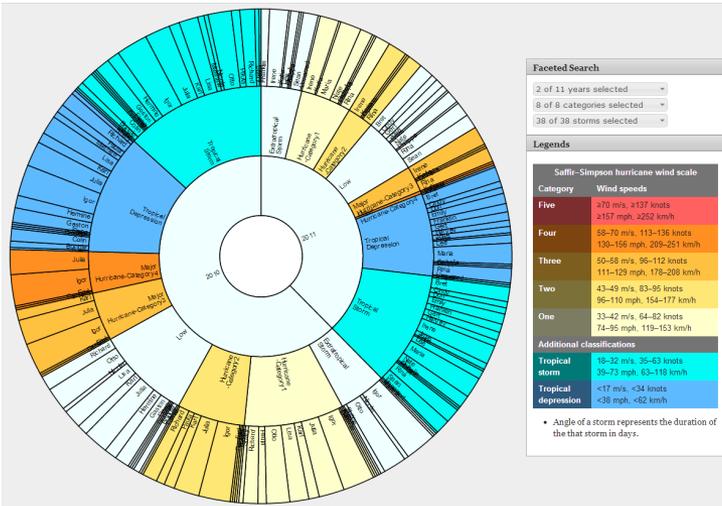
Value to Researchers?

To gather relevant data and information for case studies and climatology analysis is tedious and time consuming, current NASA Earth Science data systems are not suited to assemble multi-instrument, multi-mission datasets around specific events. Data albums addresses this need by automatically gathering all the information and data needed.

Underlying Technology?

Data Albums are auto-generated by a reusable and scalable Knowledge Synthesis Engine (KSE). KSE uses domain knowledge encoded as ontologies that captures not only what information and data resources are relevant but also the semantic relationships between the various fragments of information and data.

Visual faceted search for Hurricanes



Analytics

Box Plot Interactive Density Map Raw Data Data Summary

Interactivity

YEAR: 2001 2009 2011

Boxplot Parameters

Pressure

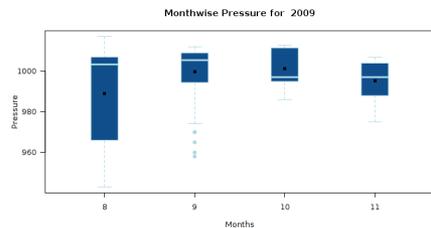
Drill-through-to-detail (Boxplot)

Drill to Months

Download Box Plot as PNG

Download Plots

Standard Boxplot with Mean



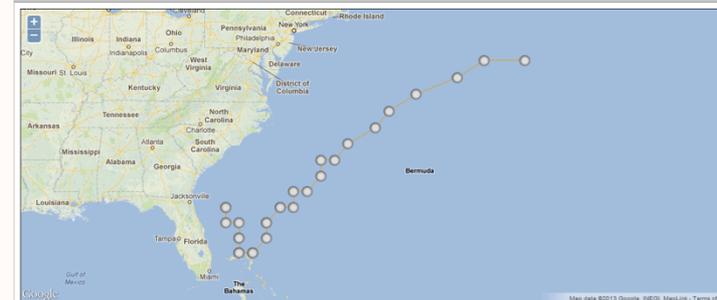
Data Albums

Compiled collections of information related to a specific science topic or event

[Back](#)

Data Album for Bret (2011)

Storm Track



Storm Statistics & Multimedia

Storm Name	Storm Year	Cyclogenesis Date	Cyclolysis Date	Highest Saffir-Simpson Category	Maximum Wind Speed	Minimum Central Pressure	ACE Index value	Damage	Track Errors	Intensity Errors	Genesis
Bret	2011	0716	0721	TS	63 mph		2.9450 x 10 ⁴	I/A			I/A

Collections based on Keywords or Instruments

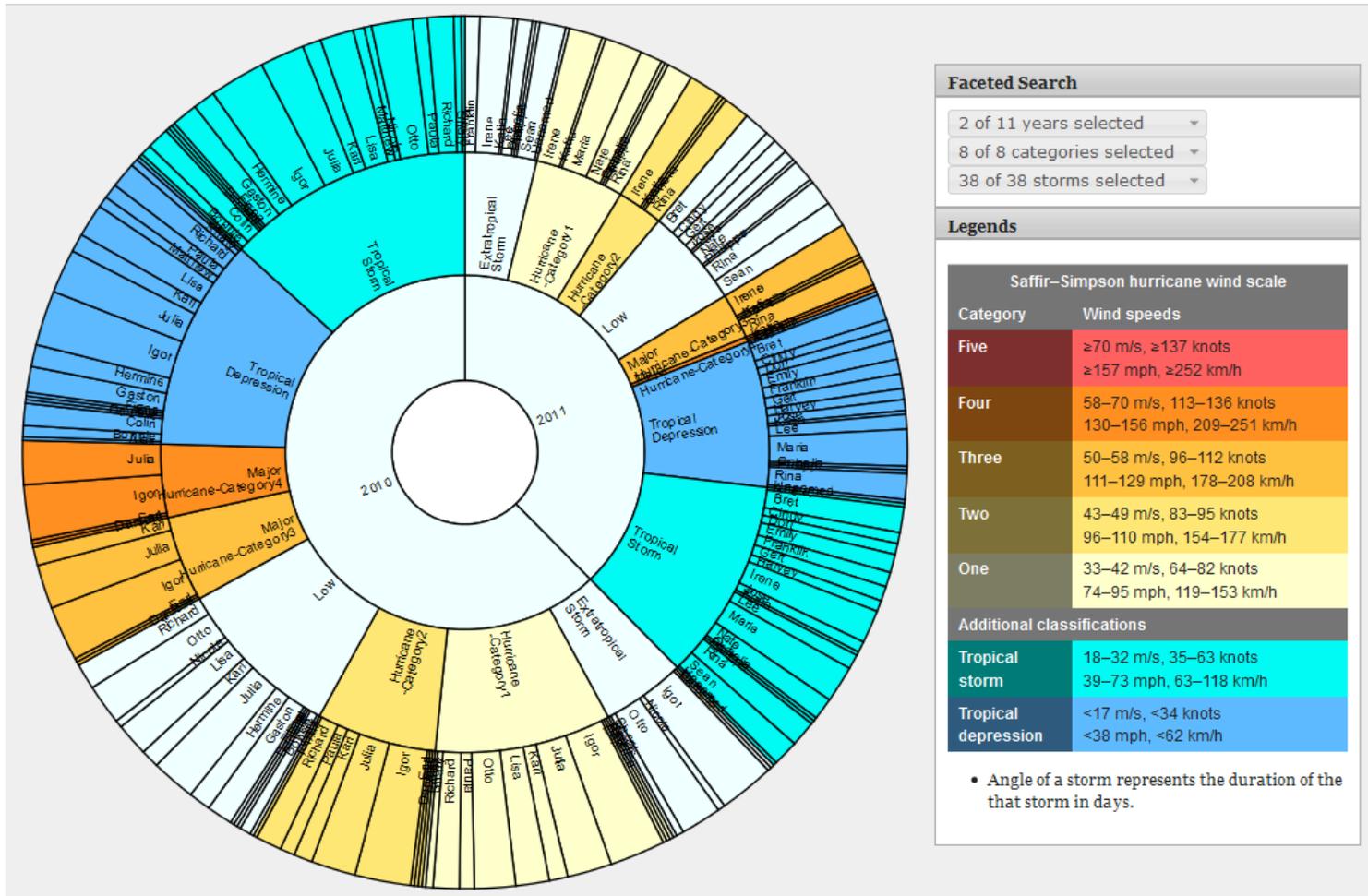


Granules

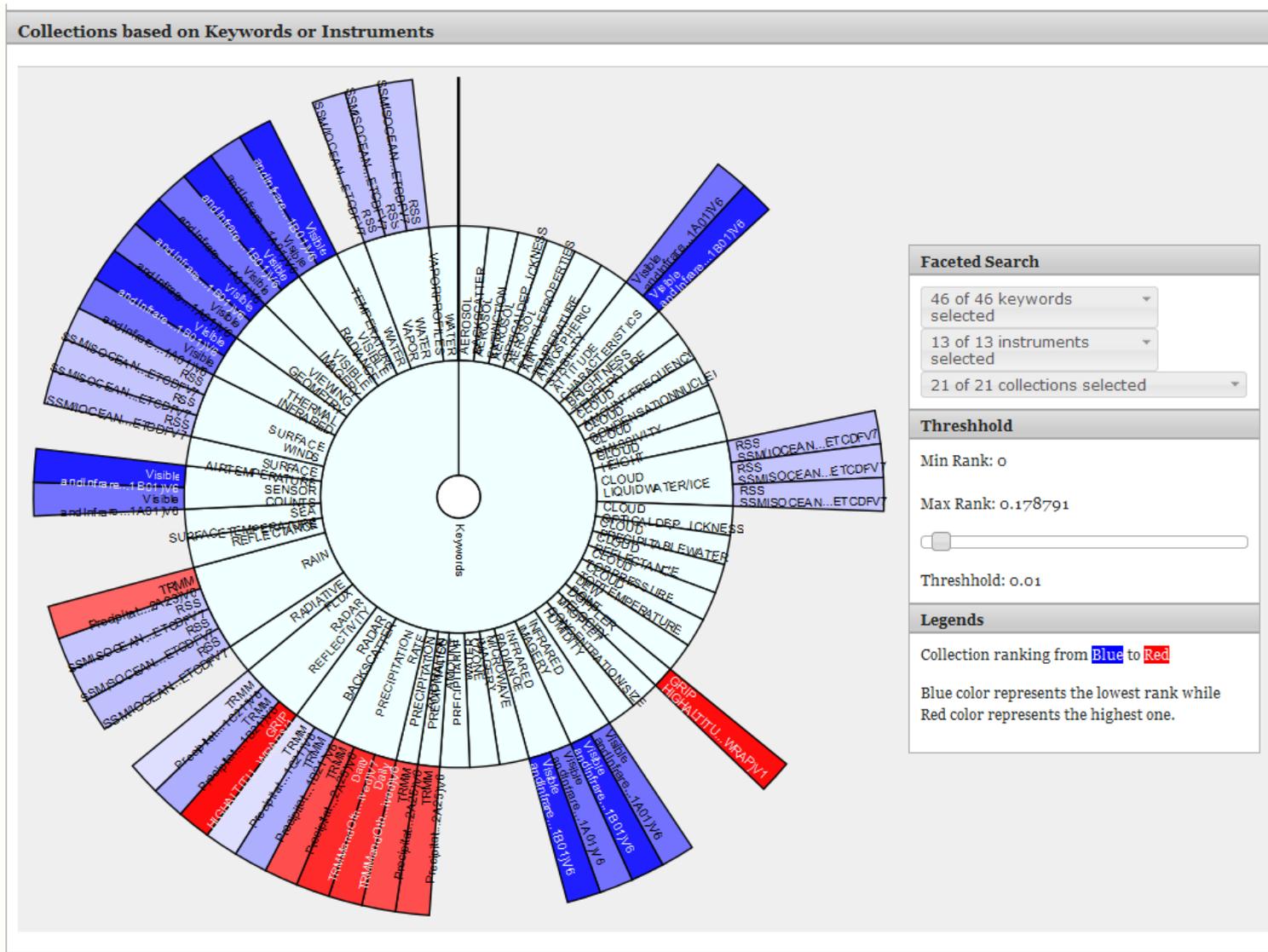
Name	Data Center	URL
3842_sally20110717.7.bin	GSFCSAPA	ftp://ds2.nascom.nasa.gov/ftp/dsta/4a4/FRMM_L3/TRMM_3842_sally20110717093842_0a
3842_sally20110718.7.bin	GSFCSAPA	ftp://ds2.nascom.nasa.gov/ftp/dsta/4a4/FRMM_L3/TRMM_3842_sally20110718093842_0a
3842_sally20110719.7.bin	GSFCSAPA	ftp://ds2.nascom.nasa.gov/ftp/dsta/4a4/FRMM_L3/TRMM_3842_sally20110719093842_0a
3842_sally20110720.7.bin	GSFCSAPA	ftp://ds2.nascom.nasa.gov/ftp/dsta/4a4/FRMM_L3/TRMM_3842_sally20110720093842_0a

[Get Links to All Granules](#) [Get Links to Selected Granules](#)

Visual Faceted Discovery Interface



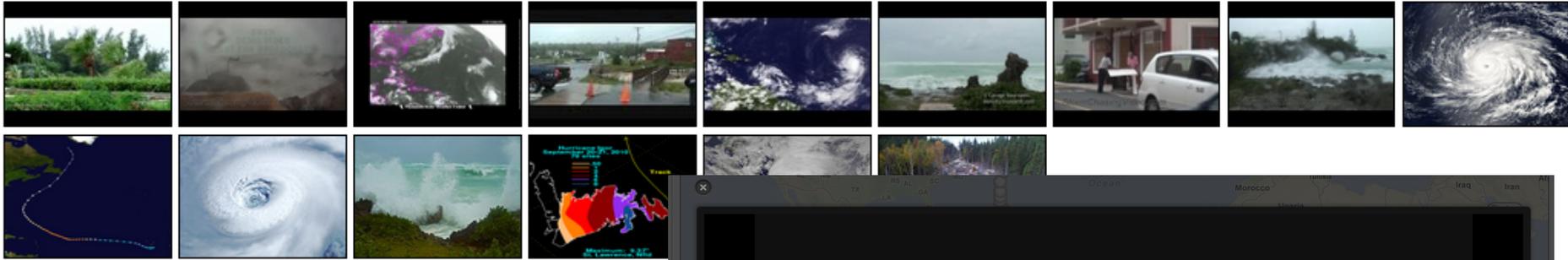
Storm View – Data Collections



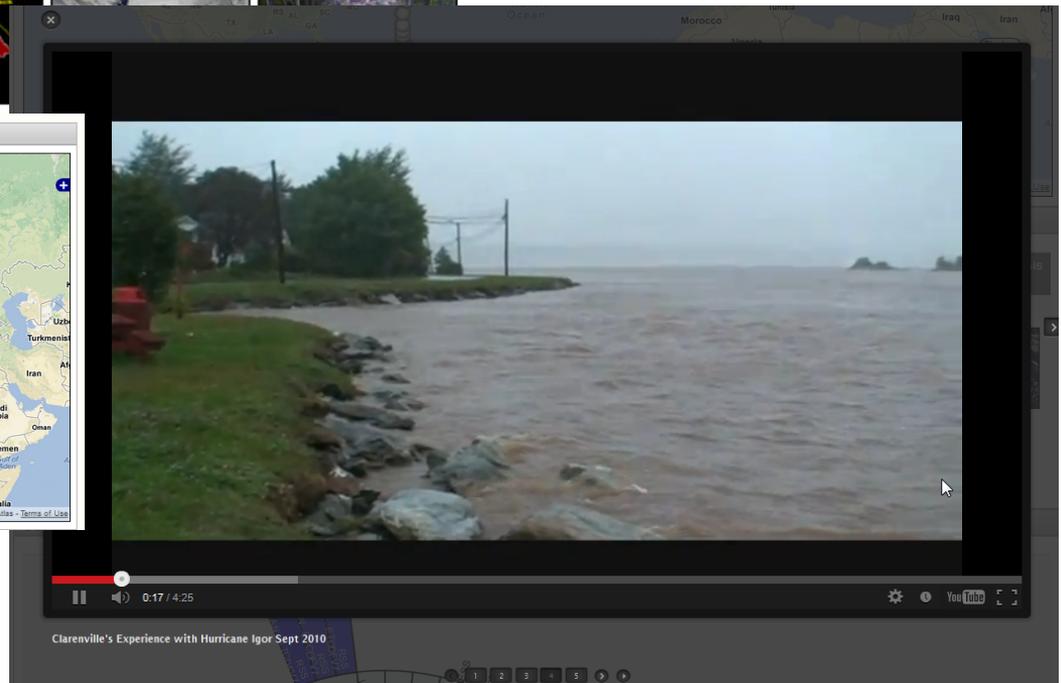
Other Aggregated Information

Storm Statistics & Multimedia

Storm Name	Storm Year	Cyclogenesis Date	Cyclolysis Date	Highest Saffir-Simpson Category	Maximum Wind Speed	Minimum Central Pressure	ACE Index value	Damage	Track Errors	Intensity Errors	Genesis
Igor	2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A			N/A



Storm Track



Summary Analytics

Interactivity

YEAR:
2001 2009 2011

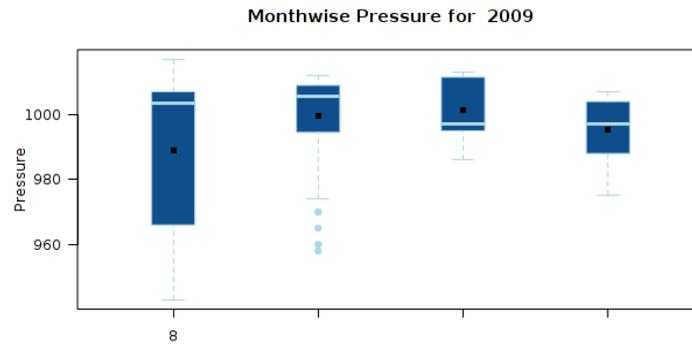
Boxplot Parameters:

- Pressure
- Maximum Sustained Wind Speed
- Storm Movement Speed
- Storm Duration

Download Box Plot as PNG

Download Plots

Standard Boxplot with Mean



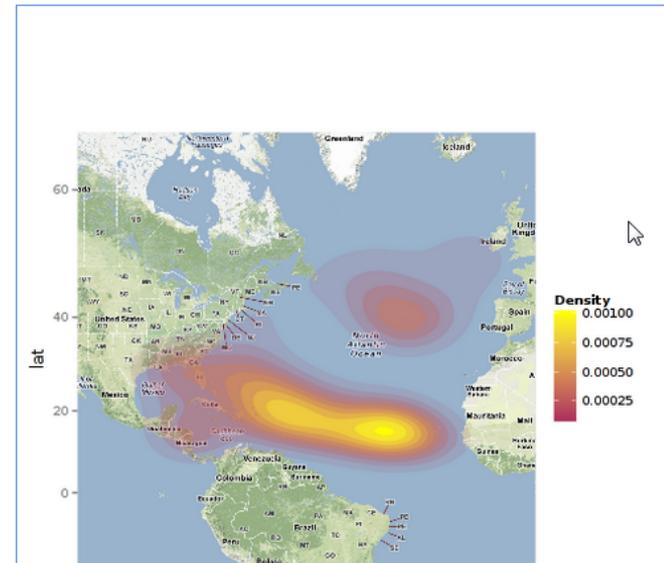
Interactivity

YEAR:
2001 2009 2011

Download Density Plot as PNG

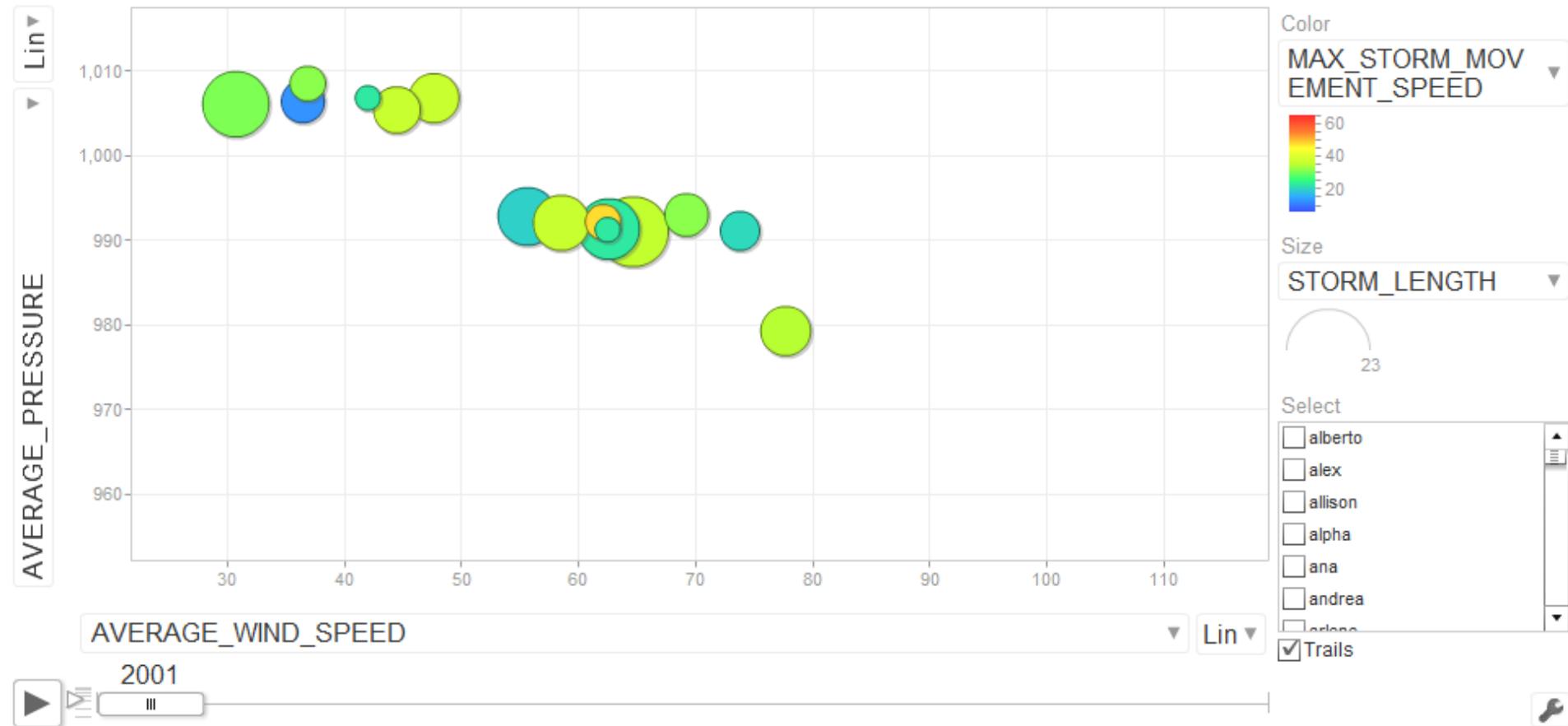
Download PNG

Event-based Density map



Interactive Analytics

Interactive Analytics



Innovations Enabled

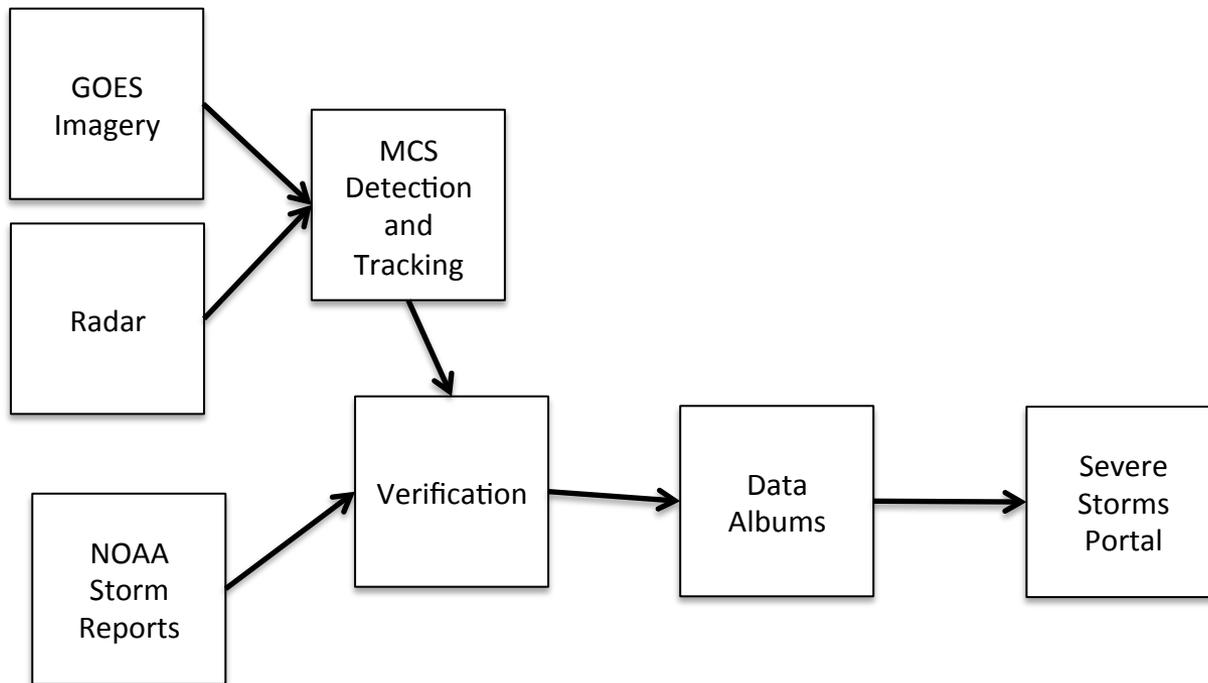
- Generalized tool designed to support “search, aggregation and access” of data and other online resources around events and *can be easily configured for other uses (disasters)*
- Automates the gathering of online resources with information filtering
- New ways of displaying search results
 - Infographics – graphic visual representation of information, data or knowledge intended to present complex information quickly and clearly
 - Results enriched with additional information (rel. score, storm cat)
- Analytics dashboard on the gathered information for events
- Use of semantic technology for relevancy ranking

Remaining Tasks

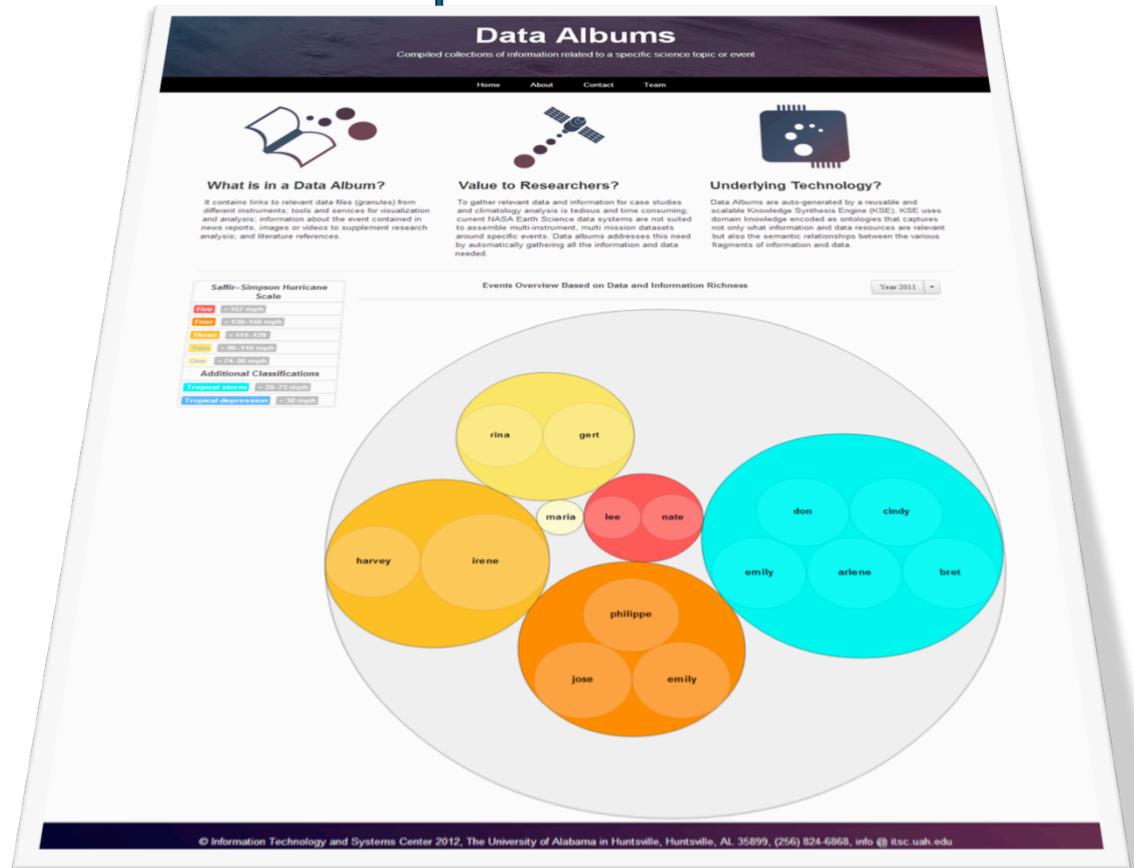
- Feature freeze and Hurricane Portal Release
 - Migrating to Dev Server, internal testing
 - Science Team Evaluation
 - Pushing to production server
- Additional Features to be added in the near future
 - Calendar View to see which data collection are available for which days
 - Displaying browse images
 - Additional data parameters for analytics (# of death, ACE index, forecast evaluations)
- SPoRT Severe storm scenario application for Year 2

Data Albums for Severe Storms

- Focus on extreme events
 - Tornadoes, Flooding, Wind Gusts, Hail Events



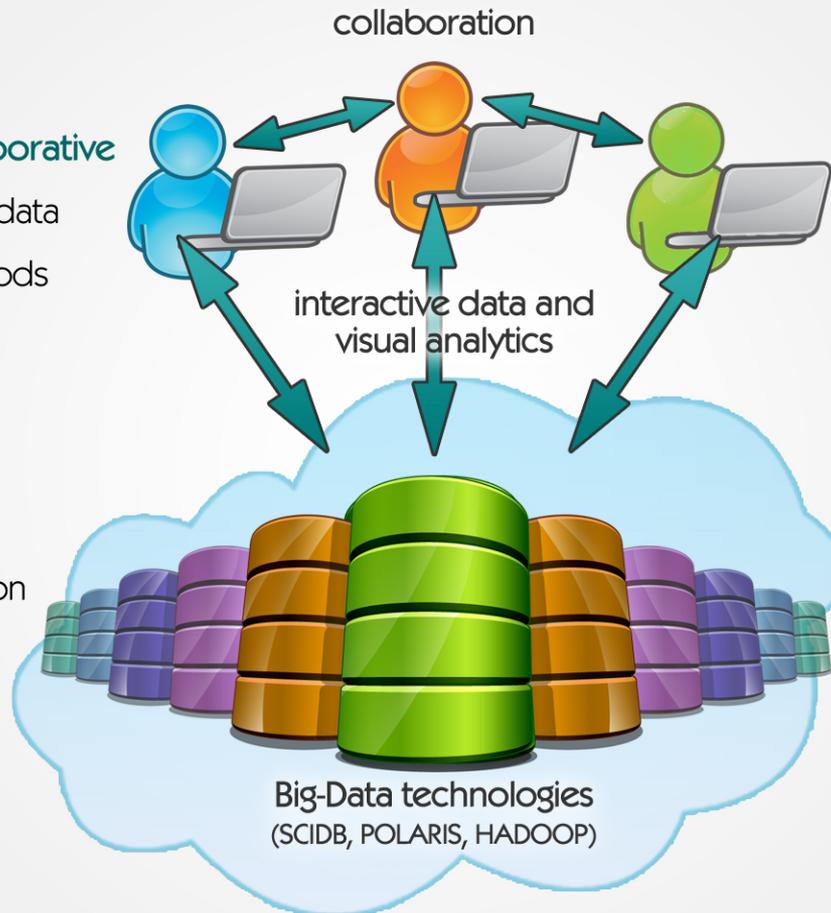
Birds Eye View of Other Projects



Automated Event Services

Utilize **Big-Data** technologies to...

- Enable **interactive** and **collaborative** scientific data analysis on big data
- Share data and analysis methods seamlessly, ...in order to...
- Relieve scientists from data management,
- Empower scientists to focus on science, and
- Boost science productivity.



- 1 Identify occurrences (events) of phenomena
 - Entities in the 4D spatiotemporal space
- 2 Associate additional relevant data with events.
- 3 Characterize phenomena with defining features extracted from data.
- 4 Correlate defining features of various phenomena in both space and time.
- 5 Improve predictions of future events using correlations among phenomena for **better decision making.**

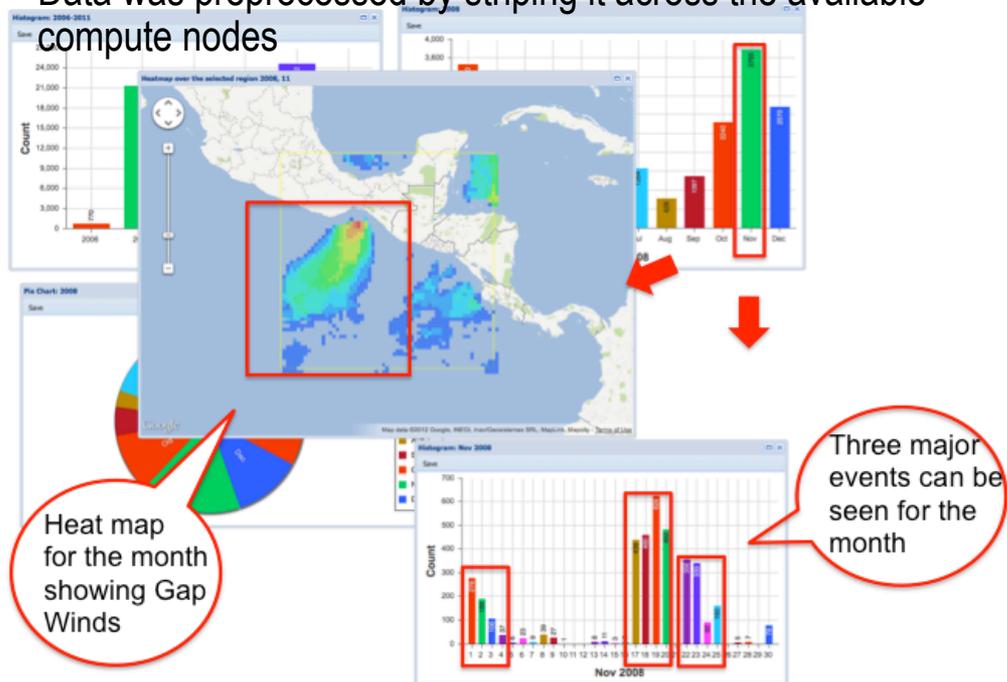
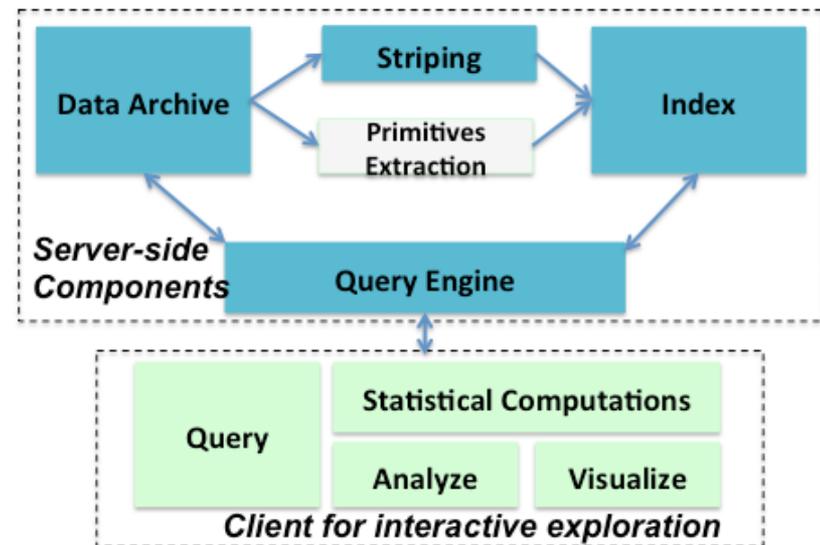
Big-Data
Vision

Technology
Infrastructure

Science
Enablement

Polaris: Discovery Engine

- All Special Sensor Microwave / Imager (/ Sounder; SSM/I and SSMIS) gridded data files of 1440x720 pixels staged on the cluster
- Data sets – Rain Rate, Surface Wind Speed, Atmospheric Water Vapor, Cloud Liquid Water
- Total volume is less than 1 Tb
- Cluster with 70 compute nodes consisting of two single core processors each used
- Data was preprocessed by striping it across the available compute nodes

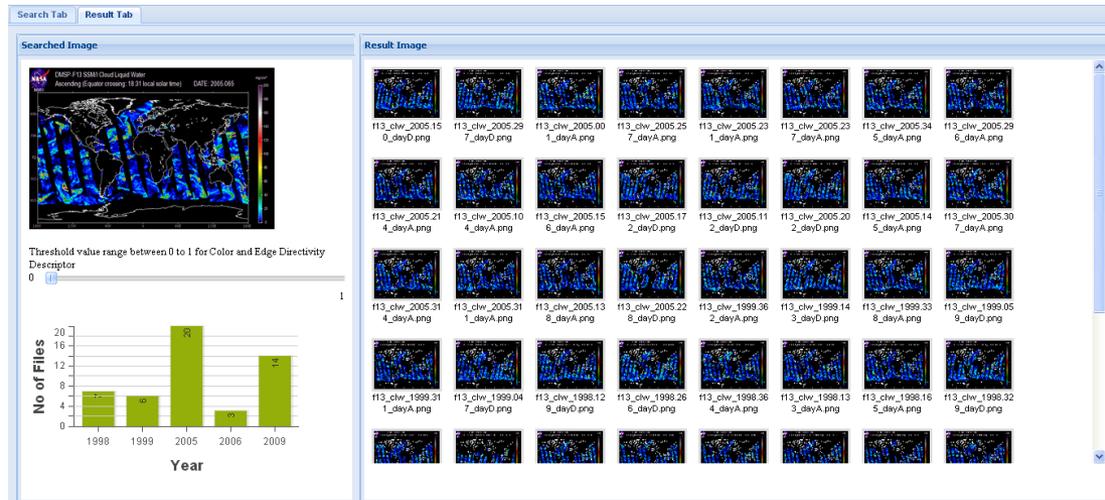
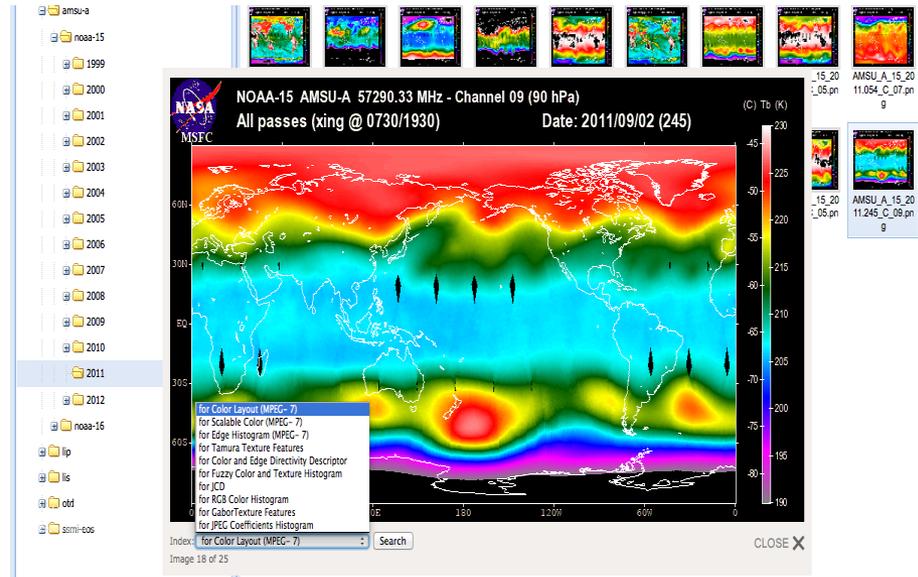


- Initial evaluations by researchers has clearly demonstrated the value of such a tool
- Polaris has been adopted by a joint project with NASA GSFC National Center for Climate Simulations
- Exploring opportunities to make the tool operational at NASA DAACs and NASA's Earth Exchange (NEX).

NASA Science on Drupal

Content Based Image Retrieval Module

- Allows users to interactively query a large database of browse images based on image contents.
- Search service provides confidence scores for the matching images and filters the images based on the scores
- Provides basic analytics on the results is available via histogram on the count of number of matching results for years, months, and days.

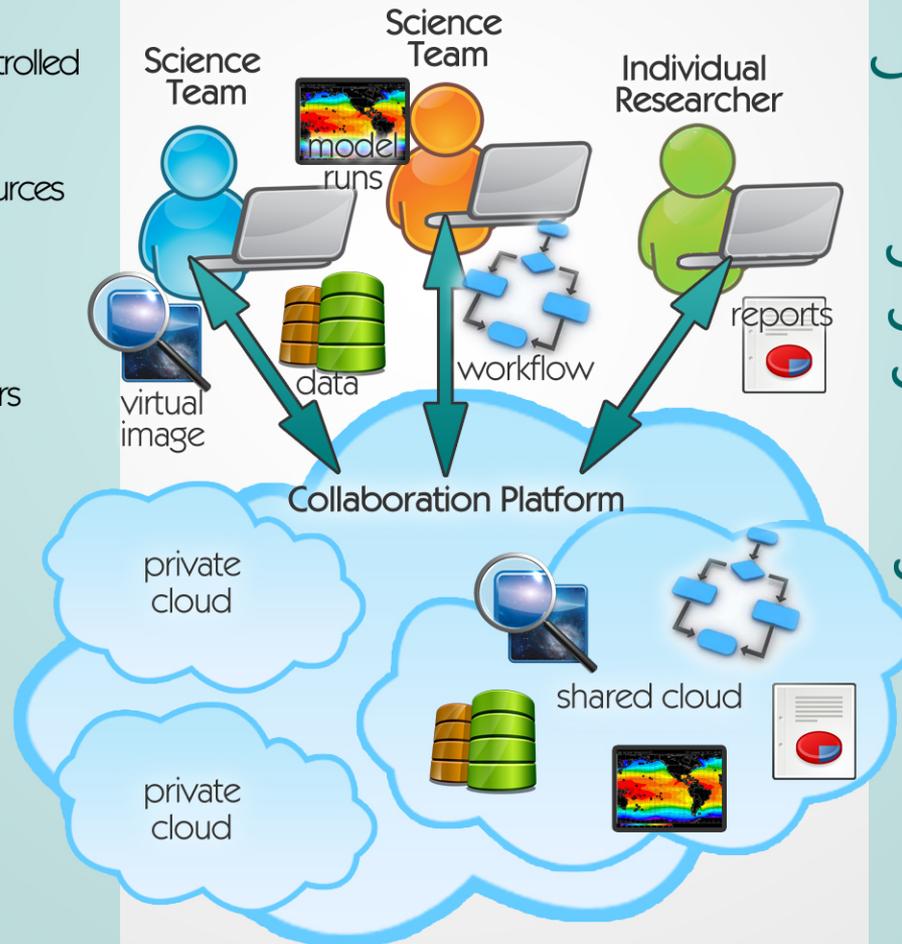


Collaborative Workbench (CWB) to Accelerate Science Algorithm Development

Sharing Knowledge is at the heart of science, yet it is challenging for researchers to effectively share information and tools

Goals

- An architecture for scalable, controlled collaboration
- Selective sharing of science resources
 - among individuals
 - within science teams
 - with the entire science community.
- Software that fits how researchers currently do scientific analysis to promote adoption



Benefits

- Accelerate science algorithm development by distributed science teams
- Reduce redundancy
- Improve productivity
- Securely share all science artifacts (data, information, workflow, virtual machines)
- Generalizable to support collaborative science algorithm development for other mission and model enterprises

Questions?

- rahul.ramachandran@uah.edu