Book Title: The Semantic Web in Earth and Space Science: Current Status and Future Directions
Publisher: IOS Press.
Editors: Peter Fox and Tom Norack
Series title: Studies on the Semantic Web
Series editor: Pascal Hitzler

Use of Semantic Technology to Create Curated Data Albums

Rahul Ramachandran[1], Ajinkya Kulkarni[2], Xiang Li[2], Roshan Sainju[2], Rohan Bakare[2], Sabin Basyal[2]

[1]NASA, Marshall Space Flight Center

[2]Information Technology and Systems Center, University of Alabama in Huntsville

Abstract

One of the continuing challenges in any Earth science investigation is the discovery and access of useful science content from the increasingly large volumes of Earth science data and related information available online. Current Earth science data systems are designed with the assumption that researchers access data primarily by instrument or geophysical parameter. Those who know exactly the data sets they need can obtain the specific files using these systems. However, in cases where researchers are interested in studying an *event of research interest*, they must manually assemble a variety of relevant data sets by searching the different distributed data systems. Consequently, there is a need to design and build specialized search and discover tools in Earth science that can filter through large volumes of distributed online data and information and only aggregate the relevant resources needed to support climatology and case studies.

This paper presents a specialized search and discovery tool that automatically creates curated Data Albums. The tool was designed to enable key elements of the search process such as dynamic interaction and sense-making. The tool supports dynamic interaction via different modes of interactivity and visual presentation of information. The compilation of information and data into a Data Album is analogous to a shoebox within the sense-making framework. This tool automates most of the tedious information/data gathering tasks for researchers. Data curation by the tool is achieved via an ontology-based, relevancy ranking algorithm that filters out non-relevant information and data. The curation enables better search results as compared to the simple keyword searches provided by existing data systems in Earth science.

1. Introduction

One of the continuing challenges in any Earth science investigation is the discovery and access of useful science content from the increasingly large volumes of Earth science data and related information available online. Common approaches used in Earth science research such as case study analysis and climatology studies involve discovering and gathering diverse data sets and information to support research goals (Shultz, 2009). Research based on case studies involves a detailed description of specific weather events using data from different sources to characterize physical processes in play for a specific event. Climatology-based research tends to focus on the representativeness of a given event by studying the characteristics and distribution of a large number of events. This allows researchers to generalize event characteristics such as spatio-temporal distribution, intensity, annual cycle, duration, etc. However, gathering relevant data and information for case studies and climatology analysis is both tedious and time consuming.

Current Earth science data systems are designed with the assumption that researchers access data primarily by instrument or geophysical parameter. Those who know exactly which data sets they need can obtain the desired files using these systems. However, in cases where researchers are interested in studying an event of research interest, they have to manually assemble a variety of relevant data sets by searching different distributed data systems. For such scenarios, the search process needs to be organized around the event rather than observing instruments. In addition, the existing data systems are based on Boolean information retrieval models utilizing keyword matches. Such systems assume users have sufficient knowledge regarding the domain vocabulary to be able to effectively utilize the search catalogs. These systems lack support for new or interdisciplinary researchers who may be unfamiliar with the domain terminology.

Consequently, there is a need to design and build specialized search and discover tools in Earth science that can filter through large volumes of distributed online data and information and only aggregate the relevant resources needed to support climatology and case studies. One specific scenario describing this need comes from hurricane science. Many hurricane researchers are familiar with limited, but specific, data sets, and are often unaware of or unfamiliar with a large

quantity of other resources. Finding airborne or satellite data relevant to a given hurricane often requires a time consuming search through web pages and data archives. Background information related to damages, deaths, and injuries requires extensive online searches for news reports and official hurricane summaries. This search process could be made much more efficient and productive if a single system could provide not just links to airborne and satellite web sites but also to specific instrument data and satellite granules relevant to the hurricane(s) as well as all related reports, summaries, news stories, and images. Not only would such a system provide a richer resource of information to the researcher, it would also increase the exposure of existing data sets to a community that may be unfamiliar with or hesitant to use the data because of difficulties in tracking down the individual satellite overpasses or airborne flights relevant to a hurricane.

This application paper presents a specialized search and discovery tool for Earth science to address these needs. The search tool automatically creates curated Data Albums. Data Albums are compiled collections of information related to a specific science topic or event, containing links to relevant data files (granules) from different instruments as well as tools and services for visualization and analysis and information about the event contained in news reports, images, or videos to supplement research analysis. Data curation by the tool is achieved via an ontology-based relevancy ranking algorithm that filters out non-relevant information and data. The search tool is novel in its use of aggregation, an ontology-based relevancy ranking algorithm to curate data, and information visualization and presentation. The key contributions of this paper are in two areas. First, the tool and its interface were designed to make the search process more intuitive and usable by supporting dynamic interaction mode and sense-making (as explained in the next section). Second, the tool uses an ontology-based relevancy ranking service for data curation, and this service provides better search results as compared to existing keyword search systems in Earth science.

The paper is structured as follows. Section 2 summarizes existing research covering the models for the search process and the use of semantic technologies in spatial data systems. The architecture design and the implementation details of specific components are covered in Section 3. The architecture and the algorithm used for the ontology-based relevancy ranking service are presented in detail in Section 3.2.3. The search tool is designed to be general and reusable. It can be customized for specific applications, and an example of this is described in Section 4 where the tool is customized to support Hurricane science. This section describes

both the application ontology needed for the customization and the design of the user interface to visually present information to support Hurricane research.

2. Background

2.1 Modeling the Search Process

Search is defined in the dictionary as trying to find something by looking or otherwise seeking carefully and thoroughly. Research on search system design has focused on the modeling "information seeking" process. The standard information seeking model used as the basis for designing most search tools breaks the process down into four steps. The first step involves identifying the problem or task. The next step entails articulating the information need. The third step is to formulate the search query, and the final step involves evaluating the results from the query. The iteration in this model involves the steps of result evaluation and query formulation. The chief drawback of this model is that it assumes that the task remains static throughout the process. Bates (1989) proposed a Dynamic Model as an alternative to the standard model. In this model, the user's needs evolve during the seeking process as he or she interacts with the retrieved information. This mode of information seeking is also referred to as "information foraging" (Pirolli, 2009). This dynamic model captures the user's "discovery of latent needs" where interaction with retrieved information leads to new and unanticipated goals.

The other component of the search relevant to the design of scientific data systems besides the process itself is sense-making. Sense-making is defined as the behavior of intelligence analysts and other knowledge workers related to information seeking and use (Dervin, 1983). Pirolli and Card (2005) define a framework for sense-making for intelligence analysts as a set of stages, with each stage utilizing the retrieved information. During the first stage (called a shoebox), the analyst gathers all relevant documents into a single collection. In the second stage, the contents of the shoebox are curated in order to be filtered further. This stage is labeled as the evidence file by Pirolli and Card. The final stage in the sense-making framework uses the information gathered in the evidence file to build a schema, a mental model of how all this information fits together.

Most Earth science data system search tools are built based on the standard information

seeking model with limited or no capability to support the aforementioned “discovery of latent needs”. The design of Data Albums supports the dynamic information seeking mode and models the sense-making framework. The auto-curated Data Albums serve as a shoebox for researchers using this tool. The contents of a Data Album can then be filtered depending on a researcher's specific need.

2.2 Use of Semantic Technologies within Spatial Data Infrastructures (SDIs)

While this list by no means is exhaustive, existing research focused on augmenting spatial data infrastructure with semantic components can be categorized into three distinct classes based on their application. These application classes are: semantic mediation and query expansion; knowledge discovery (covering search, retrieval, integration, and interpretation); and information retrieval.

2.2.1 Semantic Mediation and Query Expansion

Klien et al. (2001) used semantic technologies to address the problem of semantic heterogeneity due to synonyms and homonyms for keyword searches in catalogues. These authors used an ontology-based discovery service to mediate between the catalogs. In their prototype, each catalog has its own application ontology and the mediation is achieved by linking these application ontologies to a common shared vocabulary. Klien et al. (2001) note the limitations of such an approach, as it works in only well-defined domains where small groups commit to shared vocabulary.

As part of our earlier semantic research work with the Noesis search tool (Movva et al., 2008), we explored the use of semantic web technology for search within Geosciences. The Noesis tool was designed as a customizable search engine that uses domain knowledge captured in ontologies. The use of domain ontologies allows the richness in the relationships between the domain concepts to be captured and used by a reasoner. As advocated by Klien et al. (2001), Noesis also used mediation ontologies to mediate between the domain ontology terms (user view) and the vocabulary terms used in different resource catalogs (application view). Noesis provided a simple user interface to allow users to start the search from any point within the

ontology. If the term entered is defined in the ontology, the reasoner was used to infer all the related terms for automated search query expansion.

2.2.2 Knowledge Discovery covering search, retrieval, integration and interpretation

The use of semantic technology to support the knowledge discovery cycle within Spatial Data Infrastructures (SDIs) has been actively researched. This knowledge discovery cycle covers multiple steps including search, retrieval, data integration, and interpretation. Kammersell and Dean (2006) described the concept of "Conceptual Search" where the semantic layer is integrated within a Service Oriented Architecture. Using a prototype, the authors demonstrated the utility of the semantic layer to map user queries into component semantic queries which in turn query distributed catalogs and databases. The use case implemented by the prototype demonstrated the value of incorporating a semantic layer for data discovery and integration to answer complex geospatial queries.

In 2007, Lutz and Kolas investigated the use of semantic technology within a spatial data infrastructure to answer prototypical geospatial queries. These queries required addressing semantic heterogeneity for discovery and semantic interpretation of distributed schemas for data integration. Using a prototype, authors demonstrated the use of schema mapping rules in building descriptions of the content and schemas of the data sources in SDIs and domain rules to infer new unknown facts that can be used to answer the user's question.

Lutz et al. (2009) explored the use of semantics within a SDI at three different levels: a Metadata level focused on discovery of geographic information; a Schema level focused on retrieval; and a Data content level that addressed the interpretation, integration, and exchange of geographic information. A hybrid ontology architecture in conjunction with a DL reasoner was used to make the semantics of the information content of the geospatial web services explicit. The hybrid architecture consisted of application ontologies for each individual information systems within the SDI and a common shared vocabulary for the domain.

Peng et al. (2009) designed a prototype with a semantic middleware to automate the generation of new knowledge-added (data) products via workflows. Their prototype combines geospatial data and service discovery and process models into automatic service compositions including process modeling, process model instantiation, and workflow execution. The process models

(or “virtual data products”) in the prototype are represented via workflow ontologies, and these models link to geoprocessing workflows or service chain instances. To answer user queries, these workflows can be executed on the fly to generate required data products.

2.2.3 Information Retrieval

Semantics have been incorporated to the information retrieval systems and ranking algorithms for both improving the precision and recall of the searches and also adding additional parametric dimensions for similarity measure. For example, Liu et al. (2010) describes an approach for utilizing semantics for geographical Information retrieval that goes beyond concept mediation between a user’s cognitive model and the system and extends it to also resolve spatial relationships between objects.

Li et al. (2012) focused on developing a methodology to measure semantic similarity between spatial objects by combining Description Logic-based knowledge and a multilayer neural network. DL is used to generate features that are then used as input for the neural net.

Andrade et al. (2014) describe a framework to improve information retrievals within an SDI by extending the traditional IR concepts to include additional dimensions such as space and time in addition to themes. Similarity of a geospatial feature to the user's query is computed as a weighted sum of spatial, thematic, and temporal relevancy rank.

Our Data Albums tool utilizes semantic components within its information retrieval system and relevancy ranking algorithm. Our approach is described in detail in section 3.2.3.

3. Application Design

This section describes the Data Albums tool architecture, components, and implementation methodology. Details of the semantic component of the tool, including the ontology-based relevancy ranking service, is given in section 3.2.3.

3.1 Functional Architecture Overview

Data album creation is triggered by a real time or a historical database-driven event feed that provides spatiotemporal constraints for a given event occurrence such as a hurricane. The search tool needs to be pre-configured with a list of the different online resources it can query. Each resource has a search broker. Using this pre-configured resource list, the tool performs both data and information searches for relevant coincident observations and information. Retrieved search results are aggregated and presented to the user as an interactive Data Album containing compiled information regarding relevant data for specific events. This data is filtered based on geophysical parameters, geolocation, time, and semantic relevance. Users can review the list of data sets and, based on the information provided, obtain links to individual data files or data access services.

An application ontology describing the phenomena of interest is used by the relevancy ranking algorithm to provide a curation service to the search tool. The users can modify the semantic relevancy threshold to increase or decrease the amount of relevant data sets aggregated by the tool. The Data Albums tool also compiles supplementary information to augment research, including news articles, reports, images, and videos detailing the event and its socio-economic impacts as well as other useful information such as weather reports. The Data Albums search tool is designed using a standard thin client web server architecture as seen in Fig. 1. There are three layers of this architecture: data, service, and presentation. The data and service layers reside on the web server, whereas the presentation layer is contained within the browsers. Different components within these layers are described below.

3.2 Data Layer Components

3.2.1. Aggregation Engine

As the component responsible for aggregation, the aggregation engine invokes the different brokers that interface with the external information and data repositories. The engine can be invoked manually by the administrator or run periodically using a cron. The event broker provides all the information regarding the events, i.e., location and time. The data broker is invoked and executed using coarse grain parallelization, as the data queries for some instances

can be quite lengthy; for example, searching the NASA ECHO (NASA, 2014) catalog for Hurricanes from 1950-2012 takes around 3 months of processing on a regular web server. The NASA ECHO catalog holds information for over 6000 data collections and millions of individual data files/granules. These queries are run in the background and the tool stores the results locally.

3.2.2. Brokers

Data Search Broker - This broker is designed to query the NASA ECHO catalog. The broker uses the ECHO REST APIs (NASA, 2014) to get collections and granules. The data is retrieved for the entire duration of any given event. The data search broker uses a bounding box as a buffer for individual event location points to search for granules in the catalog. In order to make the query efficient, data granules from different collections are searched for a given time and spatial location. The broker optimizes the search queries using a temporal resolution of the data collection to avoid gathering redundant data granules within a Data Album. For example, if two points of an event are 6 hours apart, the broker should only get the granule whose time resolution covers both the time events once. These rules also make the query process more efficient as redundant queries are avoided. The data granule metadata are returned in an XML format and ingested into a MySQL database. The data search broker can be invoked on the command line and run in parallel on multiple machines to speed up the querying process. The broker also logs both download time and failures, which is useful for monitoring purposes.

Event Broker - Event information is generally located in some external database or website. A custom broker is needed to gather all the event information from these external resources. For example, the Data Albums instance for Hurricanes uses the HURDAT database (NOAA, 2014) from the National Hurricane Center. The HURDAT event broker downloads and then parses the HURDAT track information. Each point in a storm track is stored in a MySQL database. All other search brokers within the data layer use this event information to construct spatio-temporal queries.

Rule-Based Parser Broker - This broker is used to parse unstructured documents such as text files to extract information from different websites and documents. The broker uses regular expressions for parsing. Complex processing rules can be constructed by chaining a sequence

of regular expressions to match a specific pattern within the unstructured texts. These rules include options to skip spaces or ignore a certain number of characters. For example, we can use this broker to match the certain text patterns in unstructured storm reports to extract storm statistics from the NOAA website. The broker converts the extracted information into an internal information model and stores it in a MongoDB, a NoSQL database engine (MongoDB, 2014).

Another application of this broker is its use for parsing existing NOAA reports containing assessments of their ability to forecast for the genesis, intensification, and track of a hurricane. Extracting this information is important because generally researchers are interested in studying hurricanes where the forecasts were incorrect and learning why. Specific rules were defined to parse out just the key sentences in these reports.

Other Brokers - Additional brokers allow mediation between the resource API (such as YouTube and Wikipedia) and the search tool. The returned results are mapped back to our information model and stored in the MongoDB to be used by the presentation layer components.

3.2.3. Ontology-Based Relevancy Ranking Service

There are different ranking models to assess the relevancy of documents. Typical database-driven search systems use a Boolean model that only return documents containing the query keyword without any ranking. Statistical models such as *tf x idf* are the most commonly used approach to document ranking. Algorithm such as PageRank are examples of Hyperlink models. Ontology-based models are a type of Concept model that map the text in a document to concepts in an ontology via an annotation process; different algorithms can then be used to calculate weights for these annotated terms to measure the relevancy of the document.

In addition to these models, there are hybrid approaches that combine multiple models. For example, ORank (Shamsfard et al., 2006) is an ontology-based system for ranking documents that considers the conceptual, statistical, and linguistic features of a document. ORank improves precision of the search results over existing statistical model-based approaches by using concept instances in the document. Bouramoul et al. (2012) used ontologies not only for document indexing and query reformulation, but also as an important part of improving web search results by modifying information filtering and ranking algorithms. This approach also

provides a general architecture with functional components for incorporating semantics in the search process. Based on their experiments using existing search engines, Bouramoul et al. find that incorporating semantics leads to improvements in the search results.

Our approach for the relevancy ranking service utilizes a hybrid model combining statistical and concept models and is influenced by both the ORank system (Shamsfard et al., 2006) and the general architecture described in Bouramoul et al. (2012). Implemented in Java and deployed as a Tomcat Web application, this service is designed to be a general-purpose service that can be customized and re-used by other applications. The service requires an application ontology, access to collections of documents (metadata for a data set) to be ranked, and a set of theme keywords. The service then provides a list of ranked documents along with their relevancy score. For a given set of theme keywords, the service uses an algorithm that combines both ontology-based and traditional statistical scores to estimate the relevancy of a resource. The service is broken into five modules: ontology processing, information extraction, annotation, statistical analysis, and ranking. The individual components of the service and their interaction are depicted in Fig. 2.

Ontology Processing Module - This module is responsible for inference, calculating weights for individual ontological links, and calculating activation values for concepts in the ontology. The module utilizes the Protégé OWL API with the Hermit OWL Reasoner (Hermit, 2014) as the inference engine to provide relationships between all the concepts in the ontology. The ontology is treated as a graph, and weights are calculated for links between two concepts using the following equation:

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

where:

$W(c_j, c_k)$ is the weight between C_j and C_k

$\sum_{i=1}^{m} n_{i,j,k}$ is total number of related concepts to both C_j and C_k

$\sum_{i=1}^{o} n_{i,j}$ is total number of related concepts to C_j

$\sum_{i=1}^{p} n_{i,k}$ is total number of related concepts to C_k

The numerator in the formula for the weight $W(c_j, c_k)$ represents the total number of concepts in the ontology which are related to both concepts (C_j and C_k) and therefore contribute to their similarity.

The denominator in the formula scales the value of the similarity with the total number of related concepts related to each concept c_j and c_k. The obtained link weight for each concept is stored in a database.

Once the weights for the links connecting all the concepts in the ontology have been calculated, an activation value is calculated for the concepts themselves. A Spread Activation algorithm (Shamsfard et al., 2006) is used to for this purpose. For certain concepts identified as the most pertinent for thematic application, the activation value is set as one. For our application to support hurricane science, the activation value for the concept of hurricane in the ontology is set as one. The activation values for the remaining concepts within the ontology are calculated using the following formula:

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

where:

$I(c_j)$ is activation value of current concept

$I(c_i)$ is activation value of previous concept

$W(c_i, c_j)$ is link weight between two concepts

These activation values are stored in a database and used during the ranking process.

Information Extraction Parser - This parser receives a set of documents as input and extracts the text from each of them. The extracted text is stored in memory for annotation.

Annotation Module - This module is responsible for searching, parsing, and annotating the documents based on the concepts from the application ontology. The module queries the

database for stored concepts and relationships. The query returns all the terms related to the queried keyword. Each document is searched for all the concepts from the application ontology. The matching concepts from the document are saved in memory for statistical analysis.

Statistical Analysis Module - This module uses the statistic *tf-idf* formulation for each matched keyword in a document to estimate its weight. Term Frequency (*tf*) measures the frequency of each concept word (*c*) in each document (*d*).

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

where:

tf (c, d) is frequency of concept word (*c*) in document (*d*)

|*d*| is count of words in document (*d*)

Inverse Document Frequency (*idf*) measures the importance of concept word (*c*) in entire document collection (*D*):

$$idf\ (c,D) = \log(\frac{|D|}{df_{c}})$$

where:

idf (*c, D*) is inverse document frequency of (*c*) in corpus(*D*)

|*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 a matching keyword is calculated as below:

$$tf.idf(c,d,D) = tf(c,d) \times idf(c,D)$$

where:

tf.idf (*c,d,D*) is the *tf.idf* for (*c*) in (*d*) that is in corpus (*D*)

The *tf.idf* value is saved in memory and used in the ranking calculation.

Ranking Module - This module calculates a relevancy score for each document. The formula to

estimate the score (S_d) for a document (d) in the collection of documents (D) is as follows:

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

where:

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 and is calculated by multiplying the concept and statistical scores.

S_d for a given document (d) is the sum of the scores of each matched concept/keyword and is obtained by multiplying the activation values for the concept and the statistical score.

3.3 Service Layer

RESTful Data API Module - This module provides an API which supplies faceted search capabilities in the client. The API interfaces with both the MongoDB and MySQL in the data layer and enables queries to their holdings. This component also caches requests to enable efficiency and sends output responses in JSON. The JSON responses are compressed using gzip compression to enable faster transfers.

Analytics Module - This module uses the R (Bell Laboratories, 2014) statistical analysis package to generate graphs and charts. R is used in conjunction with the ShinyR and Node.js to provide the UI for the analytics module for one of the applications.

3.4 Presentation Layer

The presentation layer uses open source libraries such jQuery, jQuery UI, D3.js, OpenLayers, jqGrid, and Lightbox to create the interactive user interface and present the information visually. jQuery provides cross browser common utility functions. jQuery UI provides popup boxes, theme templates, drop down menus, sliders, and other UI elements. D3.js is a JavaScript-

based library for manipulating documents based on data (D3, 2014). D3.js provides powerful interactive visualization components using a data-driven approach to DOM manipulation while making a full use of capabilities provided by modern web browsers. Sunburst, Bubble chart, and Treemap visualizations of information are built using D3.js library. OpenLayers is used to display storm track information as well as display MODIS imagery. jqGrid (jqGrid, 2014) is used to list the available data granules for the current selection of data sets in a paged grid view UI component. Lightbox provides a slideshow plugin for viewing images and playing YouTube videos. The presentation layer accesses REST APIs provided by the service layer to fetch stored information and provide faceted search using AJAX. It is possible to build entirely new presentation layers using these REST APIs for web, desktop, or mobile clients.

4. Science Application: Catalog of Hurricane case studies

The study of an individual hurricane or a set of hurricanes requires information and scientific data related to the storm(s). We have customized the Data Albums tool to create a Hurricane Case Study portal. This portal compiles information and links to data granules from multiple instruments, organized by specific hurricane events. The portal is unique in that it supplements data set information with factual content such as news, weather reports, images, and videos, all of which are potentially extremely useful for case study analysis. The Hurricane Case Study portal uses the database of storm tracks for Atlantic and Eastern Pacific cyclones provided by the National Hurricane Center (NHC) to trigger the aggregation and curation. The portal, which is accessible via a variety of interfaces, provides a collection of data subsets and imagery of satellite observations of these hurricanes as well as data collections from NASA hurricane field campaigns including airborne observations, radiosondes, radar, and mission reports.

The presentation layer for the Hurricane portal is designed to provide an easy visual overview of all the event holdings in the catalog. The presentation layer also searches for case studies based on different feature dimensions such as storm intensity and duration and then enables a researcher to drill down to a Data Album view for a specific hurricane event. The Data Album view allows researchers to quickly assess the richness of the aggregated information and access data files needed for their investigation.

4.1 Hurricane Ontology

An application ontology is needed to customize the relevancy ranking service. Consequently, a hurricane ontology was designed with the help of hurricane researchers to capture key concepts such as when and where hurricanes occur, the different stages of hurricane life cycles and the physical parameters used to characterize them, the measurements used to study hurricanes, etc. The key concepts and relationships within this ontology are described briefly below.

Phenomena is used as the base class for all atmospheric phenomena associated with hurricanes. *Phenomena* contain several subclasses: *Cyclone*, *Flood, LandFall, Lightning, Rain, StormSurge, Tornado, TropicalDisturbance,* and *WindGust*.

The *Cyclone* class defines phenomena characterized by a rotating flow of large scale air masses; *Hurricane* is one such a phenomenon, and hence a subclass of *Cyclone.* The *Cyclone* class includes other three relevant subclasses: *Typhoon*, a similar phenomenon that occurs in West Pacific Ocean, *TropicalDepression,* and *TropicalStorm*, the latter two of which are less developed cyclone systems.

The *Parameter* class is defined as a class containing all atmospheric parameters used to characterize a hurricane. Some of the subclasses of *Parameter* include *Pressure, TemperatureProfile, RainRate, Reflectivity, WindSpeed, Velocity, StormSurgeHeight, and SurfaceTemperature*.

Component is the base class used to define all the hurricane components including *CirroStratus, Cirrus, CloudBand, Eye, CumuloNimbusCloud, EyeWall, LowPressure,* and *ThunderCloud*.

All the environmental conditions needed for hurricane formation are defined under the *Condition* base class. Example subclasses include *Coriolis*, representing coriolis force on the rotation of a hurricane; *WarmOceanSurfaceWater*, which is the warm ocean surface water needed for the formation of hurricanes; and *CyclonicRotation*, representing the cyclonic nature of hurricane morphology.

The spatial and temporal information of hurricanes is captured using *Location* and *Occurrence* base classes. Example subclasses of *Location* include *TropicalIndianOcean, TropicalNorthAtlantic,* and *TropicalWesternNorthPacific.*

Other important concepts covered in the ontology include *CasultyAndDamage*, *IntensityScale*, *DataSourceType* and *DataSourceOrigin*.

A number of object properties which relate these classes to each other are also defined. The *hasParameter* property, with its domain as the *DataSet* class and its range as the *Parameter* class, can specify the parameters that a data set contains. The *hasComponent* property relates *Component* to *Cyclone*. The *hasLocation* property has *Phenomenon* as its domain and *Location* as its range, specifying the location of phenomena. The *hasDataSourceOrigin* and *hasDataSourceType* properties are used to relate the data set source's origin and type, respectively. The *hasIntensityScale* property relates the intensity scale to a cyclone.

4.2 Visual Overview Interface

The main objective for the portal landing page (Figure 3) is to allow researchers to browse and compare different events suitable for a case study analysis or allow them to select a specific case based on year, category, and name. This view consists of the Analytics and Visual faceted search panels.

4.2.1 Analytics Panel

The Analytics panel provides users with tools to interactively and visually analyze some of the HURDAT parameters. This panel has three tabs: *Interactive, Box Plot, and Track Data*.

The *Interactive* tab displays a scatter plot which can be customized to visualize different parameters from HURDAT data - average pressure, average wind speed, maximum pressure, maximum wind speed, minimum pressure, and storm length. These parameters can be displayed in either linear or log mode by changing the axes setting of the scatterplot. Four different variables can be displayed on the scatter plot at one time. In addition to the two

parameters selected as the x and y axes, two additional parameters can be assigned to color and size. In addition, the scatter plot provides a time slider which allows the user to visually analyze different parameters for different years.

The *Box Plot* tab provides a widget to plot box plots. Box plots are commonly used to graphically view groups of data based on quartiles and are a convenient way to identify outliers in the data. Users can create box plots for pressure, maximum sustained wind speed, and duration parameters. Researchers could identify possible hurricane events for case study analysis by using these outliers. Users can also download the data regarding hurricanes, tropical storms, and tropical depression tracks and intensities from the *Track Data* tab.

4.2.2 Visual Faceted Search Panel

While the previous panel focuses on providing users with visual analytics capabilities regarding the storm track data, the Visual faceted search panel allows users to search the catalog and drill down to the Data Album for a specific storm. The panel also provides three different tabs for searching: the *Sunburst*, *Bubble Chart,* and *Classic* tabs.

The *Sunburst* tab allows a user looking for a specific storm to specify the year, category, and storm name from the pull down list. It allows the user to graphically explore the storms in the catalog. For example, after selecting specific years of interest, the user can drill down to a specific year. The Hurricanes are displayed as a wheel, arranged counter clockwise and color-coded based on their Hurricane intensity category.

The duration of these storms is represented in the angle of the wedge in the display. This interface allows a researcher to interactively browse through different storms and then find an event of interest based on category or duration suitable for their case study.

The *Bubble Chart* tab provides a different visualization of the search metadata stored in the catalog. Users can select a specific year; the storms associated with that year are presented as bubbles, the color of which represents the storm category and the size of which represents the total duration of the storm.

The Visual faceted search panel also provides the listing of the storms in a traditional tabular form in the *Classic* tab. All the storms for a given year are listed as a column, sorted alphabetically and color coded based on the storm category. This view is specifically designed to meet the needs of domain experts.

Once a user has narrowed the search down to a specific storm, double clicking on the storm name from any of the three panels launches the associated Data Album view.

4.3 *Data Album View*

The Data Album view allows the researcher to drill down for a specific hurricane event and presents all the aggregated information stored in the catalogs for the specific storm. The Data Album page has four main panels: Storm Statistics and Multimedia, Storm Track, Collections based on Keywords or Instruments, and Granules.

4.3.1 Storm Statistics and Multimedia

This panel contains a description of the storm in question gathered from Wikipedia. It also has a storm statistics table generated from parsing NOAA storm reports. The table contains useful information such as cyclogenesis and cyclosis date, highest Saffir Simpson category, maximum wind speed, minimum central pressure, damage, and errors found in genesis, track forecast, and intensity data. Youtube videos related to storm forecasts and damage reports are also presented here.

4.3.2. Storm Track

An interactive map with the storm track is displayed to provide spatial reference. Each point on the track provides the date and time of the storm at that location along with its intensity information. The map also provides two options: a basic map with a standard political map as the background, or one containing MODIS imagery as the background. The MODIS imagery is accessed in real-time using NASA's GIBS service API (NASA, 2014).

4.3.3 *Data Collections based on Keywords or Instruments*

Similar to the Visual faceted search panel on the main portal page, this panel presents the relevant data collections to the end user in any of the three views: Sunburst, Tree Map, or Classic. The user can refine his or her search process by selecting specific keywords or instruments. Selecting a keyword allows the user to quickly refine the search query and see only the curated data.

In the *Sunburst* view, the angle in a wedge for a keyword/instrument/data collection visually represents the number of data granules/files. This allows researchers to visually identify the information richness of different data collections. The *TreeMap* is generally used in information visualization to show disk usage and which directories contain the most files. It has been customized here to show the files/granules available based on keywords, instruments, or collections. The *Classic* view displays the search results in a tabular form.

The slider on the top right corner of the panel can be used to adjust the relevancy threshold. If a researcher wants to see more data collections, he can lower the relevancy threshold value. Similarly, increasing the threshold value filters and limits data collections presented in the album.

4.3.4 Granules

The results displayed on this panel are determined based on the selections made on the previous panel. This panel lists the all the granules or the data file names in a table along with relevant data center and URL links. If a user selects a specific keyword or an instrument in the previous panel, this list automatically gets filtered. Researchers can now select specific granules or all granules given in the list. Clicking on the button at the bottom of the panel generates a list of links to these granules. This list can be copied and is typically used by researchers writing their own scripts to access and analyze the data.

5. Discussion

We evaluated the value of using an ontology-based relevancy ranking service to provide data curation by comparing results against an existing search system (NASA ECHO) that supports only keyword-based searches. First we narrowed the data set listings in the NASA's ECHO catalog to focus on only one of the twelve data centers, namely Global Hydrology and Resource Center (GHRC). Using our domain knowledge and expertise, we manually identified 35 out of 160 GHRC data sets as the most relevant for studying hurricanes. The occurrence of the keyword 'hurricane' in a data set's metadata and the importance of the measured atmospheric parameters to hurricanes were some of the several criteria used in selecting these data sets. These 35 data sets serve as our truth data for evaluating the performance of the ranking service.

Searching the NASA ECHO catalog using 'hurricane' as the keyword only returned 6 GHRC data sets that were in our list of truth data, with a recall of only ~ 0.17. The evaluation result for the ranking service depends on the relevancy threshold. The relevancy threshold can be set low for lower precision and higher recall, or set high for higher precision and lower recall. For example, setting the relevancy threshold below 0.00099 returns all 160 GHRC data set with a precision of 0.21 (35/160) and recall of 1. The best combination of precision and recall is reached at relevancy threshold of 0.033. For this threshold, total of 22 GHRC data sets are returned and 19 of them match with truth data list. Hence, the service provides a precision of 0.863636 and recall is 0.542857. The relevancy threshold of 0.033 is set as the default value in the tool.

While these evaluation results show a marked improvement in search results using an ontology-based relevancy ranking service over existing traditional keyword search, there are still issues and challenges that need to be addressed. First, the existing relevancy ranking algorithm can be enhanced to improve its performance. Specifically, the existing algorithm uses the ontology primarily as a graph to calculate the concept weights and does not fully utilize the meaning of relationships between concepts described in the ontology to assign weights. Second, the performance of the relevancy ranking algorithm is dependent on the richness of the ontology. Constructing the ontology without assessing the metadata descriptions of the data sets can lead to poor performance. Finally, one of the major challenges of utilizing ontologies is the inability to scale them to cover large domains. Constructing ontologies is still a tedious manual process. For ontologies to be used in operational tools in Earth science, new methods or approaches are needed to automatically create ontologies from the existing corpus of documents. Until then,

ontologies will be used mostly in prototypes for demonstrations or in operational tools addressing niche search problems.

6. Summary and Future Work

This paper presents a specialized search and discovery tool that automatically creates curated Data Albums. The design of the tool was tailored to support the key elements within a search process, specifically enabling dynamic interaction and sense-making. The tool supports dynamic interaction with, and visual presentation of, information. The compilation of information and data into a Data Album is analogous to a shoebox within the sense-making framework. This tool automates most of the tedious information/data gathering tasks for researchers. Data curation by the tool is achieved via an ontology-based relevancy ranking algorithm that filters out non-relevant information and data. This curation enables better search results as compared to the simple keyword search provided by existing data systems in Earth science.

As part of our ongoing research and development, we have started building another instance of the Data Albums tool focusing on severe weather with a focus on selecting appropriate case studies for researchers to evaluate their numerical models. This work is a collaboration with NASA's SPoRT Center which conducts research on unique NASA products and capabilities that can be transitioned to the operational community to solve forecast problems in the 0-48 hour timeframe. SPoRT routinely runs a mesoscale configuration of the WRF model that uses NASA data sets in near real-time to evaluate the impact of NASA's satellite data in improving convective model forecasts. The severe storm instance of Data Albums will support SPoRT researchers in their model evaluation studies.

Acknowledgements

This work was funded by a NASA ACCESS Grant. The authors would like to thank Michael Goodman (NASA/HQ), Scott Braun (NASA/GSFC), Steve Berrick (NASA/HQ), and Brad Zavodsky (NASA/MSFC) for providing valuable insight during the design of the Data Albums tool.

References

Andrade, F. G., Souza Baptista, C., & Davis, C. A. (2014). Improving geographic information retrieval in spatial data infrastructures. GeoInformatica. doi:10.1007/s10707-014-0202-x

Bates, M. J. (1989). The Design of Browsing and Berrypicking Techniques for the Online Search Interface. Online Review, 13(5), 407–431.

Bouramoul, A., Kholladi, M.-K., & Doan, B.-L. (2012). An ontology-based approach for semantics ranking of the web search engines results. 2012 International Conference on Multimedia Computing and Systems, 797–802. doi:10.1109/ICMCS.2012.6320318

D3. (n.d.). D3 Javascript Library. Retrieved from http://d3js.org

Dervin, B. (1983). An overview of sense-making research: Concepts, methods and results. In Annual Meeting of the International Communication Association, Dallas, TX.

Group, I. S. (n.d.). Hermit OWL Reasoner. Retrieved from http://hermit-reasoner.com/

jqGrid. (n.d.). jqGrid. Retrieved from https://github.com/tonytomov/jqGrid

Kammersell, W., & Dean, M. (2006). Conceptual Search: Incorporating Geospatial Data into Semantic Queries. In Proceedings of Terra Cognita, Workshop of 5th International Semantic Web Conference.

Klien, E., Lutz, M., & Kuhn, W. (2001). Ontology-Based Discovery of Geographic Information Services – An Application in Disaster Management Motivating Example : Discovering Services for Estimating Storm Damage in Forests. Computers, Environment and Urban Systems, 30(1), 102–123.

Laboratories, B. (n.d.). R Langauge. Retrieved from http://www.r-project.org/

Li, W., Raskin, R., & Goodchild, M. F. (2012). Semantic similarity measurement based on knowledge mining: an artificial neural net approach. International Journal of Geographical Information Science, 26(8), 1415–1435. doi:10.1080/13658816.2011.635595

Liu, W. (2010). Ontology-based Retrieval of Geographic Information. In 18th International Conference on Geoinformatics (pp. 1–6). doi:10.1109/GEOINFORMATICS.2010.5567612

Lutz, M., & Kolas, D. (2007). Rule-Based Discovery in Spatial Data Infrastructure. Transactions in GIS, 11(3), 317–336. doi:10.1111/j.1467-9671.2007.01048.x

Lutz, M., Sprado, J., Klien, E., Schubert, C., & Christ, I. (2009). Overcoming semantic heterogeneity in spatial data infrastructures. Computers & Geosciences, 35(4), 739–752. doi:10.1016/j.cageo.2007.09.017

MongoDb. (n.d.). MongoDb. Retrieved from https://www.mongodb.org/

Movva, S., Ramachandran, R., Graves, S., & Conover, H. (2008). Customizable Search Engine with Semantic and Resource Aggregation Capability. In The Semantic Web meets the Deep Web Workshop, IEEE Joint Conference on E-Commerce Technology and Enterprise Computing, E-Commerce and E-Services . Washington DC.

Nance, C., Losser, T., Iype, R., & Harmon, G. (2013). NoSQL vs RDBMS-Why There is Room for Both. In In Proceedings of the Southern Association for Information Systems Conference (pp. 111–116).

NASA. (n.d.-a). Global Imagery Browse Service. Retrieved from https://earthdata.nasa.gov/about-eosdis/system-description/global-imagery-browse-services-gibs

NASA. (n.d.-b). NASA ECHO API.

NASA. (n.d.-c). NASA ECHO Catalog. Retrieved from https://earthdata.nasa.gov/echo

NOAA. (n.d.). Hurricane database. Retrieved from http://www.nhc.noaa.gov/data/#hurdat

Pirolli, P. (2009). An elementary social information foraging model. In Proceedings of the 27th international conference on Human factors in computing systems (pp. 605–614).

Pirolli, P., & Card, S. (2005). The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis 3333 Coyote Hill Road 2 . A Notional Model of Analyst Sense- making. In Proceedings of the 2005 International Conference on Intelligence Analysis, Mclean, VA. Mclean, VA.

Schultz, D. M. (2009). Eloquent Science: A Practical Guide to Becoming a Better Writer, Speaker, and Atmospheric Scientist (p. 440). AMS Books.

Yue, P., Gong, J., Di, L., He, L., & Wei, Y. (2009). Integrating semantic web technologies and geospatial catalog services for geospatial information discovery and processing in cyberinfrastructure. GeoInformatica (Vol. 15, pp. 273–303). doi:10.1007/s10707-009-0096-1

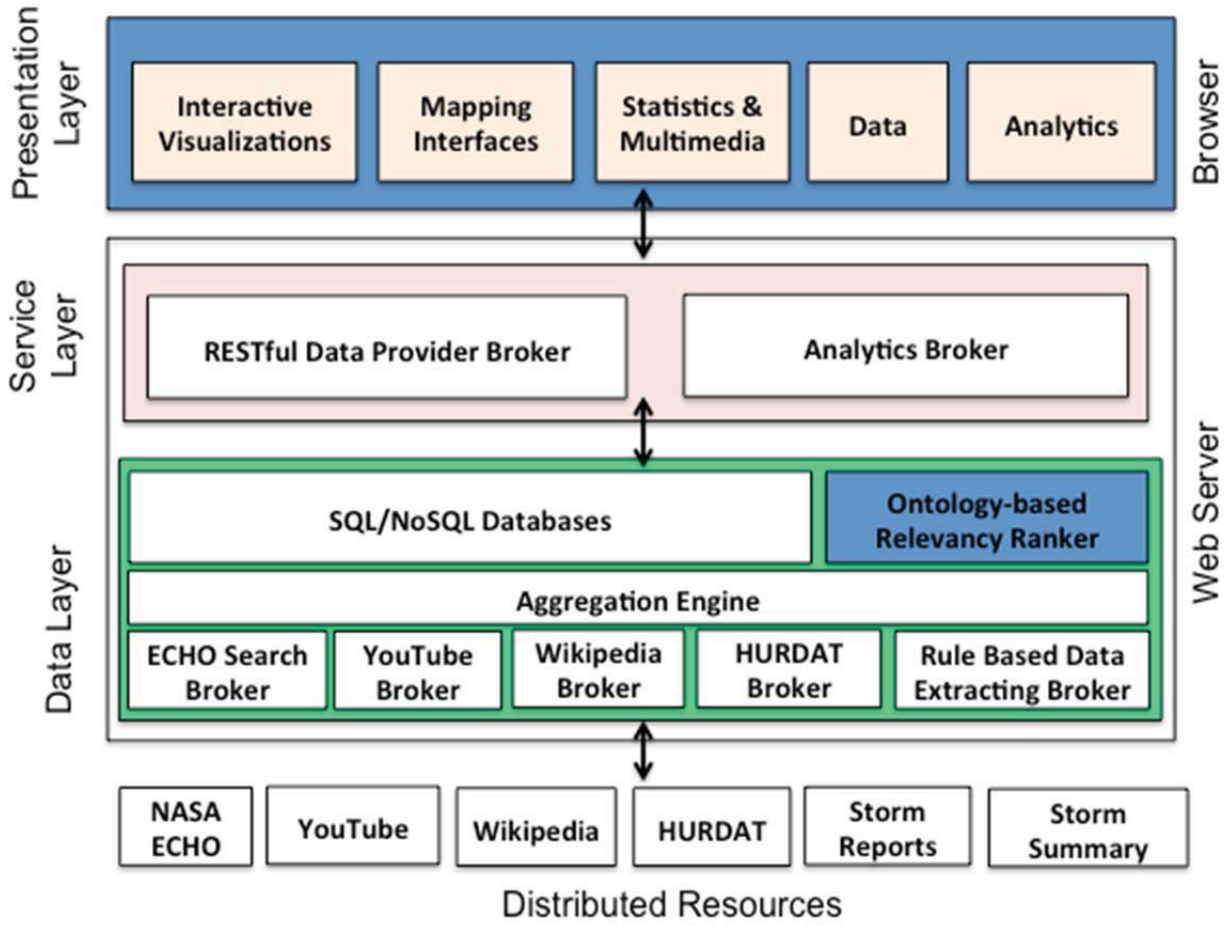

Figure 1: Client Server architecture for the Data Album tool depicting different layers and functional components

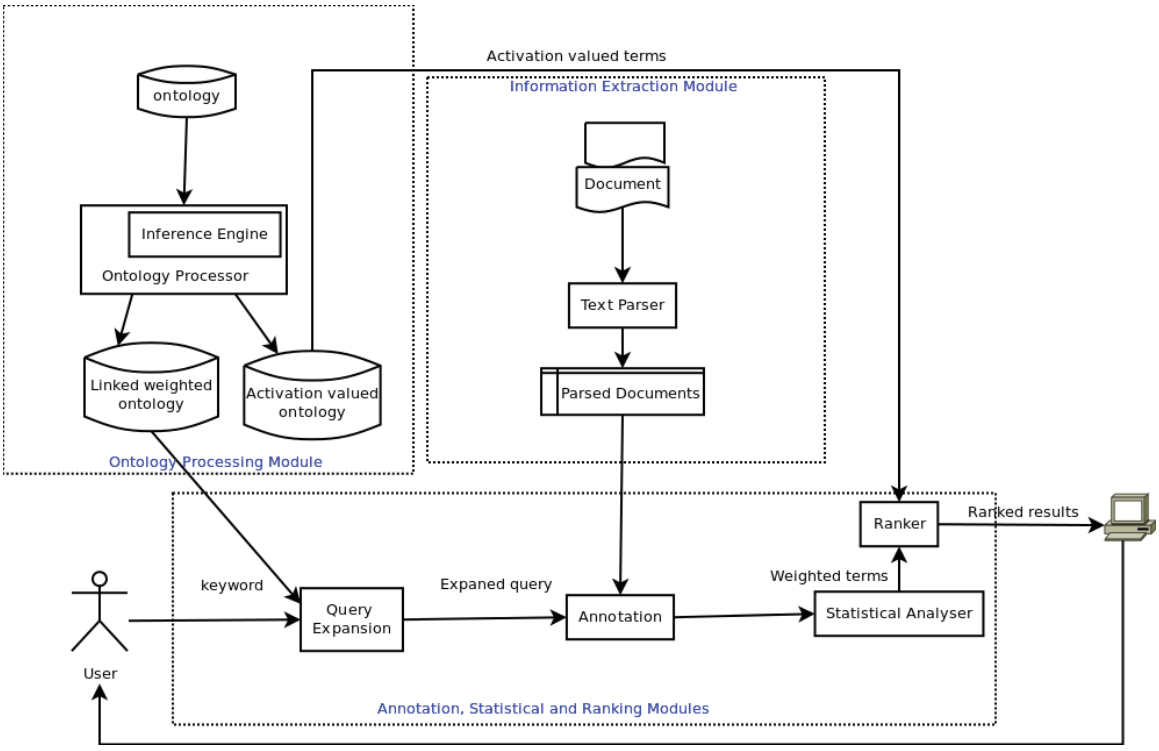

Figure 2: Ontology-based relevancy ranking service components and their functional interactions.

Figure 3. The main page for the Hurricane portal built using the Data Albums tool. The main page provides researchers a high level view of all the events and allows them to interact with and analyze the curated catalog holdings.

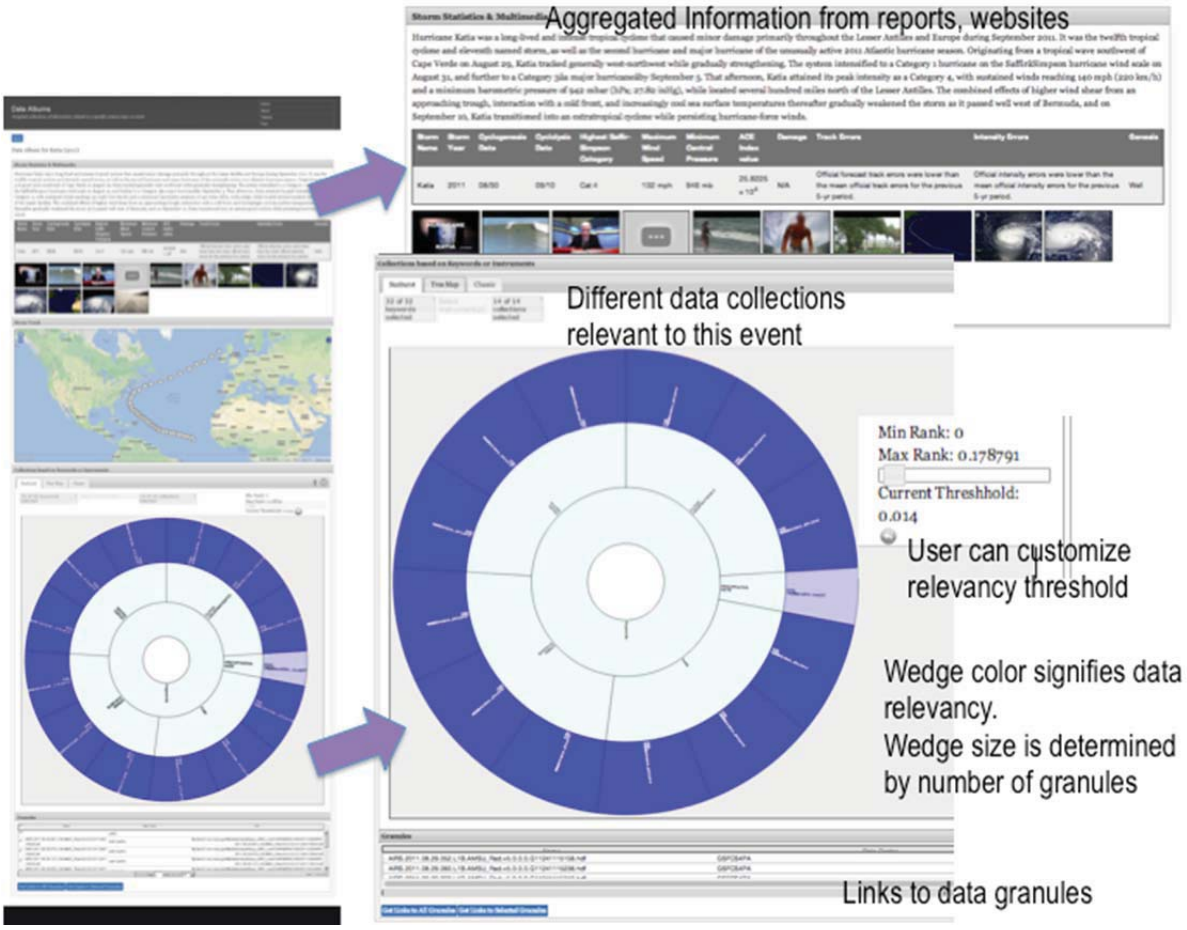

Figure 4: The Data Album view provides the researchers the deep dive functionality where they can look at all the aggregated information and curated data gathered by the tool for a specific hurricane event, assess data richness, and drill down to individual files need to support their study.