View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by NASA Technical Reports Server

DATA ALBUMS: *AN EVENT DRIVEN SEARCH, AGGREGATION AND CURATION TOOL FOR EARTH SCIENCE*

Rahul Ramachandran[1], *Ajinkya Kulkarni*[2], *Manil Maskey*[2], *Rohan Bakare*[2], *Sabin Basyal*[2], *Xiang Li*[2] *and Shannon Flynn*[2]

[1]NASA/MSFC
[2]University of Alabama in Huntsville

1. INTRODUCTION

One of the largest 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. Approaches used in Earth science research such as case study analysis and climatology studies involve gathering discovering and gathering diverse data sets and information to support the research goals. 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 characteristics such as spatio-temporal distribution, intensity, annual cycle, duration, etc.

To gather 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 the datasets of interest can obtain the specific files they need using these systems. However, in cases where researchers are interested in studying a significant event, they have to manually assemble a variety of datasets relevant to it by searching the different distributed data systems. In these cases, a search process needs to be organized around the event rather than observing instruments. In addition, the existing data systems assume users have sufficient knowledge regarding the domain vocabulary to be able to effectively utilize their catalogs. These systems do not support new or interdisciplinary researchers who may be unfamiliar with the domain terminology.

This paper presents a specialized search, aggregation and curation tool for Earth science to address these existing challenges. The search tool automatically creates curated “Data Albums”, aggregated collections of information related to a specific science topic or event, containing links to relevant data files (granules) from different instruments; tools and services for visualization and analysis; and information about the event contained in news reports, images or videos to supplement research analysis. Curation in the tool is driven via an ontology based relevancy ranking algorithm to filter out non-relevant information and data.

2. DATA ALBUMS TOOL: SYSTEM DESIGN

2.1 Architecture Overview

Data Album creation is triggered by event "feed" such as an RSS news feed that specifies spatiotemporal constraints. The search tool needs to be pre-configured to list of 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, filtered based on geophysical parameters, geolocation, time and scientific relevance. Users can review the list of datasets, and based on the information provided, obtain links to individual data files or data access services. An ontology based relevancy ranking algorithm provides a curation service that is used by the search tool. The users can modify the 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, and other useful information such as weather reports. All entries in a Data Album include links to the original source.

The Data Albums search tools is designed using a standard thin client-web server architecture [1]. There are three layers - data, service and presentation. The data and the service layer reside on the web server whereas the presentation layer is contained within the browsers.

2.2 Data Layer Components

Key component responsible for the aggregation, *BuildDB* engine within the Data Layer invokes all the different brokers that interface with all the external information and data repositories. The engine can be invoked manually by the administrator or run periodically using a cron. *Data broker* is invoked and executed using coarse grain parallelization as the data searches for some instances can last for a long time. The data is retrieved for the entire duration for any given event. The 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 uses rules to such as checking to see the temporal resolution of the data collection to avoid gathering redundant data granules within a data album. Event information is generally located in some external database or website. A custom *event broker* is needed to gather all the event information from these resources. The spatio-temporal information gathered by the event broker is used by all other search brokers within the data layers. A *Rule Based Parser broker* is used to parse unstructured documents to extract relevant information from different websites and documents. The broker uses regular expression for parsing, and complex rules can be created to chain a sequence of regular expression to match a specific pattern. Other brokers such as YouTube and Wikipedia, allow the search tool to mediate between the resource API and the search tool. All the results returned from these brokers is mapped back to an internal information model and stored in a database. *Ontology Based Relevancy Ranking Service* implemented in Java (and deployed as Tomcat Web App) is designed to be general-purpose service that can be customized and re-used in many different applications. The relevancy ranking service requires an application ontology, access to collection of documents to be ranked and set important keywords. Given a set of keywords, the service uses an algorithm that combines both ontology based and traditional statistical scores [2,3] to estimate relevancy of a resource against the application ontology. The service then returns a sorted list of ranked documents along with their relevancy score.

2.3 Service Layer

RESTful Data API [4] Module component provides an API to enable faceted search capabilities in the client. The API interfaces with the databases in the data layer and enables queries on their holdings. This component also caches requests to enable efficiency and sends the output responses in JSON. The JSON responses are compressed using gzip compression to enable faster transfers. The Analytics Module uses R statistical analysis package to generate graphs and charts. R is used in conjunction with the ShinyR and node.js to provide the UI for analytics.

2.4 Presentation Layer

Presentation layer uses open source libraries such jQuery, jQuery UI, D3.js, OpenLayers, jqGrid and Lightbox. Presentation layer accesses REST APIs provided by Service Layer to fetch storm information and provide faceted search using AJAX. The flexible architecture design enables one to build entirely new presentation layers using these REST APIs for web, desktop or mobile clients.

3. SCIENCE APPLICATION: CATALOG OF HURRICANE CASE STUDIES

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 backed by a rich catalog of case studies. Unlike the current hurricane portals that focus on either visualization or specific datasets, this portal compiles information about and direct links to data granules from multiple instruments, organized by specific hurricane events. The portal is unique in that it supplements the dataset information with factual content such as news, weather reports, images, and videos 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), accessible via a variety of interfaces; a collection of data subsets and imagery of satellite observations of these storms; and data collections from NASA hurricane field campaigns, including airborne observations, radiosondes, radar, and mission reports.

A snapshot of the Hurricane Portal is presented in Fig 1 below.

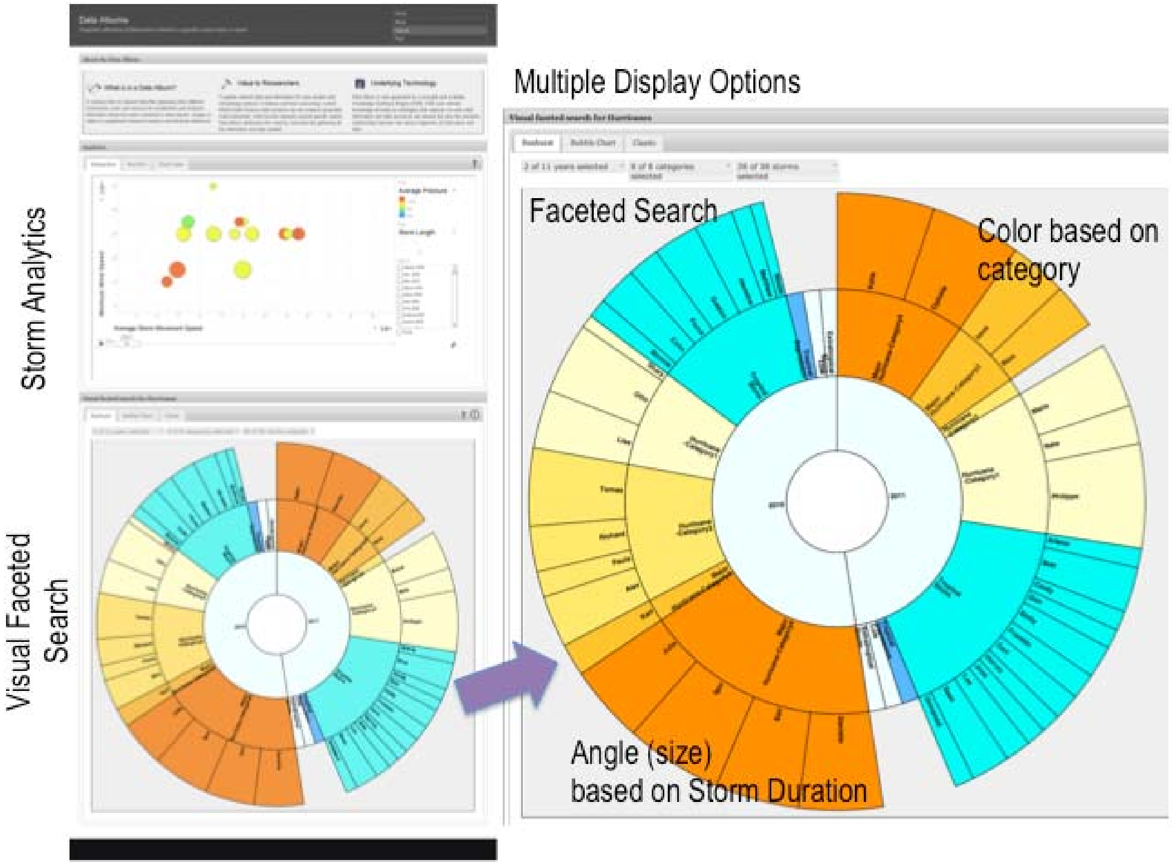

Fig. 1. Snapshots from the main page of the Hurricane Portal showing the Storm Analytics panel and the Visual Faceted search functionality. Users can drill down to a specific storm (event) using this visual interface to get to the data album for a specific event.

4. REFERENCES

[1] http://msdn.microsoft.com/en-us/library/ee658117.aspx

[2] A. Bouramoul and M. Kholladi, “An ontology-based approach for semantic ranking of the web search engines results,” in 2012 International Conference on Multimedia Computing and Systems (ICMCS), 2012.

[3] 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.

[4] Fielding, Roy Thomas (2000), Architectural Styles and the Design of Network-based Software Architectures, Doctoral dissertation, University of California, Irvine

[5] http://wwwdev.itsc.uah.edu/dataalbums