January 2010

Integration Of Multi-Sensory Earth Observations For Characterization Of Air Quality Events

Erin Robinson
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS
School of Engineering and Applied Science
Department of Energy, Environmental and Chemical Engineering

Thesis Examination Committee:
Rudolf Husar, Chair
Pratim Biswas
Jay Turner

INTEGRATION OF MULTI-SENSORY EARTH OBSERVATIONS FOR CHARACTERIZATION OF AIR QUALITY EVENTS

by

Erin Marie Robinson

A thesis presented to the School of Engineering
of Washington University in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

December 2010
Saint Louis, Missouri

ABSTRACT OF THE THESIS

Integration of Multi-Sensory Earth Observations

for Characterization of Air Quality Events

by

Erin Marie Robinson

Master of Science in Energy, Environmental and Chemical Engineering

Washington University in St. Louis, 2010

Research Advisor: Professor Rudolf Husar

In order to characterize air quality events, such as dust storms or smoke events from fires, a wide variety of Earth observations are needed from satellites, surface monitors and models. Traditionally, the burden of data access and processing was placed on the data user. These challenges of finding, accessing and merging data are overcome through the principles of Service Oriented Architecture. This thesis describes the collaborative, service-oriented approach now available for air quality event analysis, where datasets are turned into services that can be accessed by tools through standard queries. This thesis extends AQ event evidence to include photos, videos and personal observations gathered from social media websites such as Flickr, Twitter and YouTube. In this thesis, the service-oriented approach is demonstrated using two case studies. The first explains the benefits of data reuse in real-time event analysis focusing on the 2009 Southern California Smoke event. The second case study highlights post-event analysis for EPA's Exceptional Event Rule. The thesis concludes with a first attempt to quantify the

benefits of data reuse by identifying all of the different user requirements for Earth observation data. We found that the real-time and post-event analysis had 68 unique Earth observation requirements making it an ideal example for illustrating the benefits of service oriented architecture for air quality analysis. While this thesis focuses on the air quality domain, the tools and methods can be applied to any area that needs distributed data.

Acknowledgment

The author would like to thank Dr. Rudy Husar for his guidance and support over the last eight years. She would also like to thank Kari Hoijarvi and Ed Fialkowski for their programming expertise, without these two, her thesis would not have been possible.

The author would like to acknowledge the funding sources that have supported this work NASA, EPA and ESIP.

Finally, the author would like to thank R.J. Sak and her family for their endless support as she has pursued her dreams.

Erin Marie Robinson

Washington University in St. Louis
December 2010

Table of Contents

Abstract ii

Acknowledgment iv

List of Figures vi

1 Introduction 1

2 Traditional Air Quality Event Characterization 3
2.1 April 2003 Kansas Smoke Characterization 3
2.2 Traditional Data Access 5

3 Service-Oriented Air Quality Analysis 7
3.1 DataFed 9
3.1.1 DataFed Wrappers 9
3.1.2 DataFed Metadata 10
3.1.3 DataFed Workflow 11
3.2 Social Media as an Air Quality Sensor 13
3.3 Combining Science Data and Social Media for AQ Event Analysis 18

4 Case Study: Real-time Event Analysis of August 2009 Southern California Fires 20

5 Case Study: Exceptional Event Analysis for May 2007 Georgia Swamp Fires 23
5.1 May 2007 Georgia Swamp Fires 24
5.1.1 A: Event Identification 25
5.1.2 B: Clear Causal Relationship between the Data and the Event 27
5.1.3 C: The Event is in Excess of the "Normal" Values 28
5.1.4 D. The Exceedance or Violation would not Occur, But For the Exceptional Event 30
5.1.5 Summary 31

6 Earth Observations Requirements 32

7 Future Work 35
7.1 Collaboration on AQ Event Analysis 35
7.2 Global Earth Observing System of Systems 35

References 37

Vita 37

List of Figures

Figure 2.1 Kansas smoke event, April 12 2003. a.Fire pixels from MODIS satellite sensor, b. fine particle mass concentration from FRM c. organic fine particle mass concentration from VIEWS, d. noon total reflectance and e., f. the aerosol optical thickness derived from the SeaWiFS satellite sensor. 5
Figure 2.2 Schematics of traditional client-server 6
Figure.3.1 Schematics of OGC standard protocols, WMS and WCS. 7
Figure 3.2 Service Oriented approach where many data providers are loosely connected to many user applications 8
Figure 3.3 Federated vs. stovepipe data system architecture 9
Figure 3.4 AQ uFIND – DataFed Catalog for finding and accessing data. 11
Figure 3.5 A typical web service program for the creation of multi-layer data views. 12
Figure 3.6 DataFed Browser for browsing data in space, time and parameter 13
Figure 3.7 YouTube, Flickr and Twitter displaying material published to the web during 2009 SoCal Fires.... 15
Figure 3.8 a. Air Twitter Information System b. AirTwitter timeseries 17
Figure 4.1 Air Quality EventSpace for 2009 Southern California Wildfires.... 21
Figure 4.2 a. Entire site traffic for ESIP wiki Aug. 16-Sept.16; b. Traffic to SoCal EventSpace driven from Twitter; c. Geographic location of traffic. 22
Figure 5.1 a. GA Smoke, b. Fire pixels c. Airnow surface $PM_{2.5}$ and OMI Absorbing Index 24
Figure 5.2 FRM data layer. Red circles show magnitude of surface concentration in exceedance 25
Figure 5.3 Analyst console showing a variety of data to provide spatial context for the event.... 26
Figure 5.4 Social Media available to document the event.... 26
Figure 5.5 Trajectory analysis around Okefenokee Fire 27
Figure 5.6 VIEWS a. Sulfate and b. Organic Carbon $PM_{2.5}$ speciation; NAAPS c. Sulfate and d. Smoke 28
Figure 5.7 Three year time series for a site impacted by GA Smoke. Yellow indicates the +/- 15 day window.... 28
Figure 5.8 Anomaly detection a. actual measurement b. 84^{th} percentile from 3 years c. difference.... 29
Figure 5.9 Areas identified for Exceptional Event status 30
Figure 6.1 Earth Observation by EE Rule Step 33
Figure 6.2 EO Requirement by User Type 34
Figure 7.1 Global Earth Observing System of System (GEOSS, 2005) 36

Chapter 1

Introduction

Traditionally, air quality analysis was a slow, deliberate investigative process occurring months or years after the monitoring data had been collected (Husar and Poirot, 2005). Satellites, real-time pollution detection and the World Wide Web have changed all that. The era of agile air quality analysis began in the late 1990s, when real-time satellite images became available through the Internet. High-resolution color satellite images were uniquely suited for early identification and tracking of extreme natural or anthropogenic aerosol events. In April 1998, for example, a group of analysts keenly followed and documented in real-time the trans-continental transport and impact of Asian dust from the Gobi desert on the air quality over the Western US. (Husar, et. al. . , 2001). Soon after, in May 1998, a well-documented incursion of Central American forest fire smoke caused record fine mass ($PM_{2.5}$) concentrations over much of the Eastern US (Peppler, et. al. . , 2000). During such extreme air quality events, managers need 'just in time analysis', not just air quality data. Real-time analysis can explain the causes, the status and the likely evolution of the events.

While technological advances have been made recently, many have stated the need for improving information systems to improve analysis and ultimately provide more societal benefits. The Decadal Survey (Anthes and Charo, 2005) from the National Research Council emphasizes the need for improving information systems used for decision support. In order to do improve these systems, the participating organizations will need to make a dramatic shift from traditional emphasis on self-reliance toward more collaborative operations — a shift that will allow the community as a whole to perform routinely at levels unachievable in the past (McConnell, 2008).

The instantaneous 'horizontal' diffusion of information via the Internet now permits, in principle, the delivery of the right information to the right people at the right place and time.

Standardized computer-computer communication languages and Service-Oriented Architectures (SOA) now facilitate the flexible processing of raw data into high-grade 'actionable' knowledge. Last but not least, the Web has opened the way to generous sharing of data and tools and faster knowledge creation through collaborative analysis and virtual workgroups.

This thesis will describe the collaborative, service-oriented approach now available for air quality (AQ) event analysis. The thesis focuses on the air quality domain, but the same tools and methods could be used for other regulatory needs, informing the public or research applications.

Chapter 2

Traditional Air Quality Event Characterization

The characterization of gaseous species requires only the physical dimensions of space and time (x,y,z,t). The full characterization of the particulate air pollution system requires four additional dimensions: aerosol chemical composition, C, particle size, D, particle shape, S and nature of aerosol mixing, X (x, y, z, t, D, C, S, X). A single monitor can only measure a subset of these dimensions. For instance, satellite sensors have high spatial resolution (x,y) but they detect the radiative effect of a vertical column. Hence, a satellite sensor detects an aggregate contribution, which is an integral over five dimensions (z, D, C, S, X). On the other hand, surface monitors have high temporal and composition resolution, but lack any information to the spatial distribution. Combined the two sensors can give an idea of elevation of the aerosol, and if on the surface the composition as well as the spatial distribution (Husar, 2010). Consequently, a full spatial, chemical and optical characterization of the air pollution system requires the combined use of many pollutant sensors.

2.1 April 2003 Kansas Smoke Characterization

Smoke events are some of the most visible air quality events, both on the surface, impacting daily life, and from a satellite perspective. These events have unique composition and depending on the size of the fire, the smoke can spread over a region far larger than the area actually burned. In order to characterize the smoke impact, both surface and satellite monitors are needed.

Fortunately, both the air quality monitoring and data dissemination technologies have also advanced considerably since the 1990s. Recent developments offer outstanding

opportunities to fulfill the information needs for the new agile air quality management approach. The data from surface-based air pollution monitoring networks now provides routinely high grade, spatio-temporal and chemical patterns throughout the US for $PM_{2.5}$ and ozone. Satellite sensors with global coverage and kilometer-scale spatial resolution now provide real-time snapshots, which depict the pattern of haze, smoke and dust in stunning detail. The 'terabytes' of data from these surface and remote sensors can now be stored, processed and delivered in near real time. Air quality analysts can now observe air pollution events as they unfold, 'congregate' over the Internet in ad hoc virtual work-groups to share their observations and collectively gather the insights needed to explain the observed phenomena.

For several days every spring, the refuse from the agricultural fields in Kansas-Oklahoma are burned resulting in major smoke plumes that cover multi-state areas of the Midwest. The fire pixels (Fig. 2.1a), obtained from satellite and other observations, provide the most direct evidence for the existence and location of major fires. In Figure 2.1a, the cluster of fires in Kansas is evident.

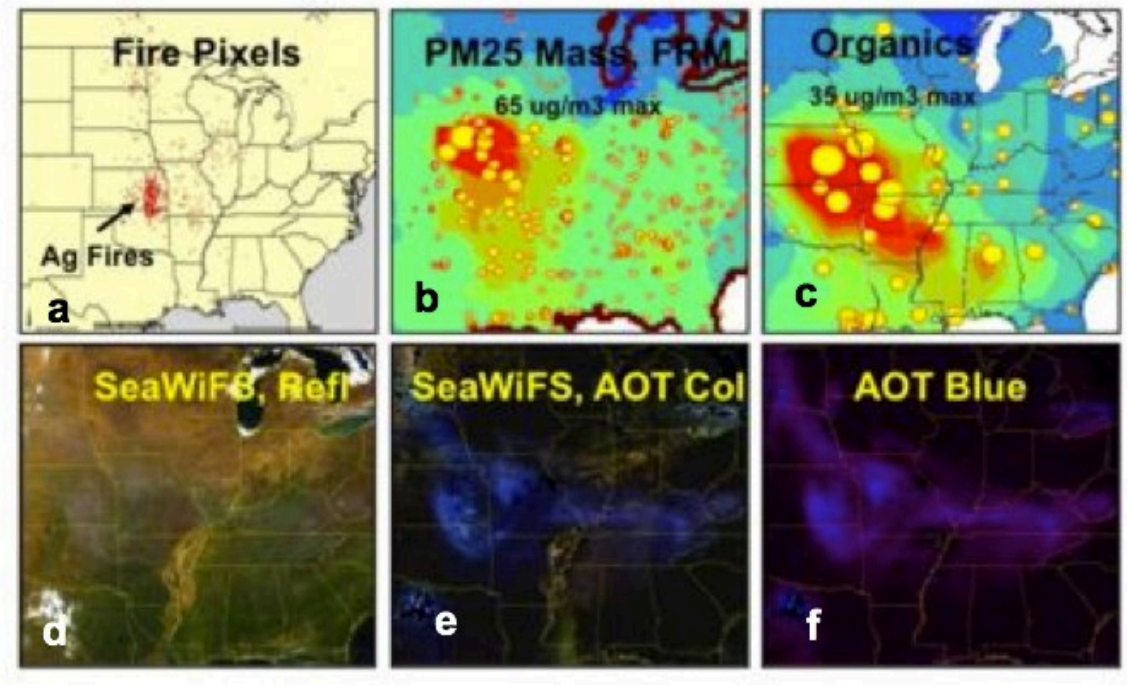


Figure 2.1 Kansas smoke event, April 12 2003. a.Fire pixels from MODIS satellite sensor, b. fine particle mass concentration from FRM c. organic fine particle mass concentration from VIEWS, d. noon total reflectance and e., f. the aerosol optical thickness derived from the SeaWiFS satellite sensor.

The spatial pattern of $PM_{2.5}$ (Fig. 2.1b) comes from the EPA Federal Reference Method monitoring network. The FRM network establishes the spatial and temporal patterns of the key pollutants $PM_{2.5}$ and ozone. Figure 2.1c shows the spatial pattern of organic fine mass which is part of the chemical speciation data from VIEWS data system[i]. Together these surface monitors identify that there is aerosol on the surface and that it is comprised of organics.

The true color SEAWiFS satellite image (Fig. 2.1d) show a blueish haze over Kansas. The Aerosol Optical Thickness (AOT), derived from SEAWiFS (Raffuse, 2002), show a strong, back-scattering signal in the blue wavelength as well, further confirming the presence of smoke (Fig. 2.1e,f). Together these datasets confirm that on April 12, 2003 the haze over the Midwest was due to smoke and where the organic $PM_{2.5}$ was high, the smoke was impacting surface air quality.

2.2 Traditional Data Access

No single dataset in the above Kansas Smoke example characterizes the spatio-temporal distribution of the smoke. For this characterization, four different datasets were used: two satellite datasets and two surface monitors. None of these datasets have the sole purpose of characterizing smoke, however because they are accessible via the web, they were brought together to describe the event.

Even if all the data were accessible through the Internet, researchers would go to every data provider to access and download the data and then run that data through specialized, "stovepipe", processing routines for their purpose. In the client-server architecture (Fig. 2.2) the individual servers are designed and maintained as autonomous systems, each delivering information in its own way.

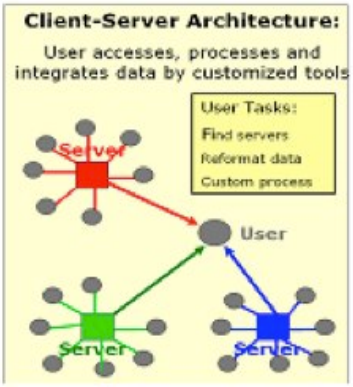


Figure 2.2 Schematics of traditional client-server relationship

For example, SEAWiFS raw data are stored within the NASA DAAC as swaths, the satellite data still warped the way that the satellite passed over. To create the images seen in Figure 1d, we ordered satellite swaths covering a geo-rectangle of interest, received an e-mail when the package was ready, downloaded the data and unzipped it two times (due to the size). Once the raw data were on our computers, we then preprocessed the swaths to georeference the satellite data, Rayleigh correct the images and splice them together to form a single image for the U.S. We then further processed the data in order to extract AOT (Fig. 2.1e,f). Similarly, each of the other datasets was accessed, downloaded and used through a multi-step process.

The traditional process described above is tedious and places the burden of data use on the data user. There are many hurdles to overcome from the user perspective. She doesn't know a certain type of data exists, if she does then she can't access, if she can access then she isn't sure about the quality and if she does find that is good quality, she can't merge it with other datasets (NAS, 1989).

Chapter 3

Service-Oriented Air Quality Analysis

These challenges of finding, accessing and merging data are overcome through the principles of Service Oriented Architecture (SOA). While catalogs or brokers have been around for a long time, the key benefit of the SOA approach is the loose coupling between service providers and service users. Loose coupling is accomplished through plug-and-play connectivity facilitated by standards-based data access service protocols. SOA is the only architecture that we are aware of that allows both loose, dynamic connection and seamless flow of data between a rich set of provider resources and diverse array of users.

The dynamic connection or data federation is accomplished by turning data stored and exposed through a server into a data service. Data as a service makes it accessible to other computers through standard query interfaces and communication protocols (Fig. 3.1).

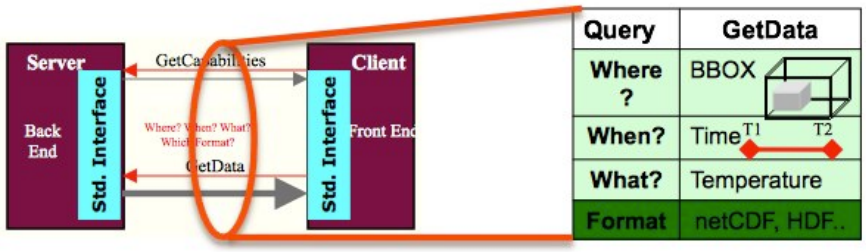


Figure.3.1 Schematics of OGC standard protocols, WMS and WCS.

The air quality community has adopted the Open Geospatial Consortium (OGC)[ii] Web Map Services (WMS) and Web Coverage Services (WCS) as the standard query interface for requesting and delivering air quality (AQ) data.

Data providers "publish" data in a catalog, users "find" data in the catalog and when ready, they connect or "bind" to the selected data access service. Users of the federated data can then access the federated resource pool through suitable catalogs. From the user's

perspective, federating the data makes the physical location irrelevant. This loosely-coupled networked architecture is consistent with the "publish-find-bind" triad of SOA (Fig. 3.2). The result is a dynamic binding mechanism for the construction of loosely-coupled work-flow applications.

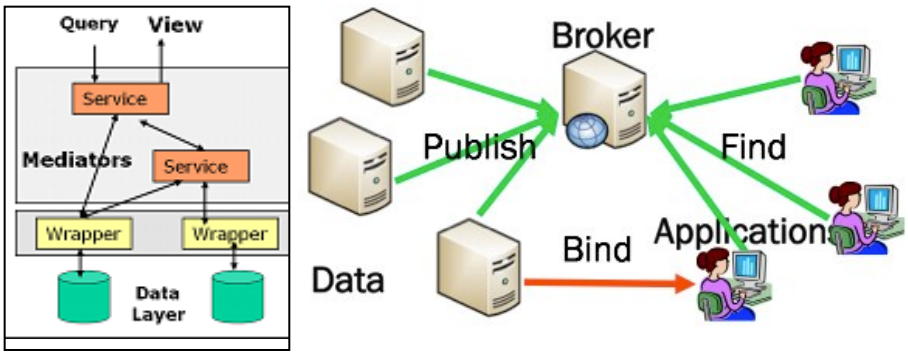


Figure 3.2 Service Oriented approach where many data providers are loosely connected to many user applications

This viewpoint of the information system architecture describes the functional relationship between the distributed components and their interaction at the interfaces. This viewpoint highlights the key difference between the traditional client-server architecture and the loosely coupled, networked architecture, where queries and views are mediated by services (Fig. 3.2). Service interfaces accomplish the chores of homogenizing the distributed, heterogeneous datasets allowing the user to access data from multiple servers and immediately use their standards-based tools speeding up the analysis process from weeks or months to days or weeks.

Service Oriented Architecture has been accepted as the desired way of delivering Earth Observation data products. However, the adoption of formal standards-based data-as-service offerings has been slow within NASA, NOAA, EPA and other Agencies. Offering images through OGC WMS standard interface is becoming common for many Federal Agency data products, but there is currently no effective way for the users to find those services since they are dispersed over many web pages. As a consequence, current SOA information systems are rather fragile. Autonomous data access and processing services can be integrated into application software for data exploration and analysis using appropriate workflow software.

3.1 DataFed

The federated data system, DataFed[iii] (Husar and Poirot, 2005), is a mediated integrator of heterogeneous, distributed data (Fig. 3.3).

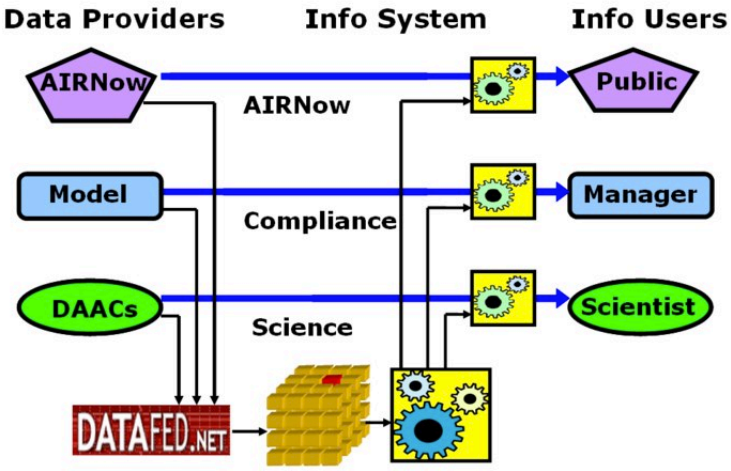


Figure 3.3 Federated vs. stovepipe data system architecture

The architecture of DataFed is consistent with the SOA approach of creating all components as reusable services. For data providers, DataFed offers wrappers to turn heterogeneous data into data services and then provides a catalog for the end user to find and browse metadata for registered datasets. Applications that access standard data services and process the data using workflow chains have also been developed as a part of DataFed.

3.1.1 DataFed Wrappers

Many historic datasets don't have standard interfaces built-in. Wrappers provide uniform interfaces to heterogeneous data by compensating for physical access and syntactic differences (Husar and Poirot, 2005). Each wrapper has two sides, one facing the heterogeneous data source that requires custom programming. Data wrappers incorporate the physical server location, perform the space-time subsetting services, execute format translations etc. The other side of the wrapper faces outward toward the internet cloud and

presents the uniform interface to the heterogeneous data, i.e. turning data into machine-consumable services.

The wrapper can be physically located on the same server as the data source. However, in a networked environment, such as DataFed, the wrapping process can be performed as a service by a third party. The placement of third-party wrapper components between network nodes is desirable for all network links, not only for legacy connections. They allow modification of service connections in response to environmental changes, e.g. an update of an interface standard without intruding on the operation of the data server. The result of this 'wrapping' process is an array of homogeneous, virtual datasets that can be queried by spatial and temporal attributes and processed into higher-grade data products.

3.1.2 DataFed Metadata

In current catalogs, metadata mainly covers finding and accessing data since the provider or distributor of the data provides metadata. This metadata includes intrinsic discovery metadata such as spatial and temporal extent, keywords and contact information for the provider. Metadata also includes distribution information for data access.

In DataFed, the user experience is improved by adding additional AQ-relevant metadata added to provider-contributed metadata in order for the AQ user to easily find the data. This additional metadata allows for sharp queries to be given in the parameter, time, and physical space. Another feature of the user-centric system is that using web analytics additional metadata are attached to each dataset in order to provide information about dataset usage characteristics. DataFed has its own catalog, AQ uFIND[iv], where data access services can be registered for standards-based access for processing, visualization and exploration.

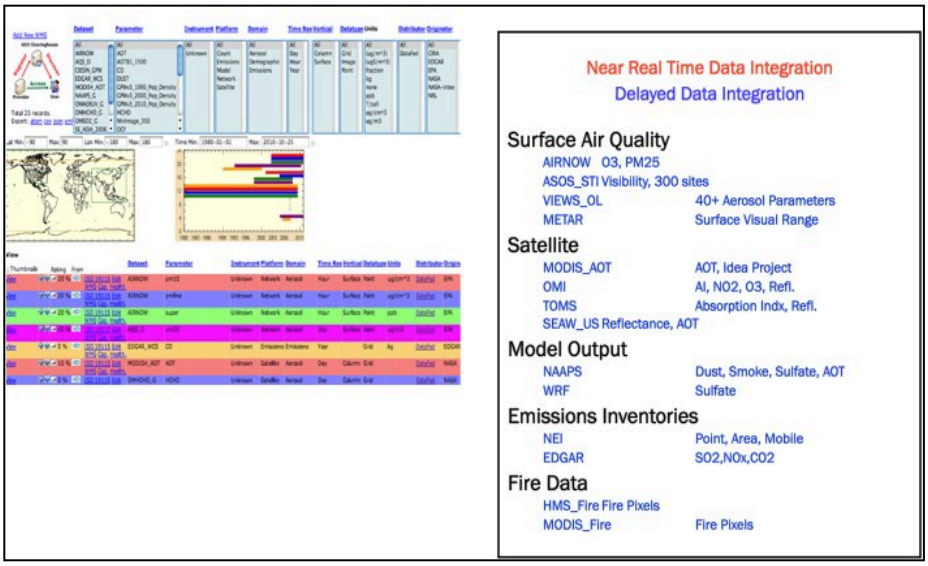


Figure 3.4 AQ uFIND – DataFed Catalog for finding and accessing data.

AQ uFIND is the web service-based tool set: **U**ser-oriented **F**iltering and **I**dentification of **N**etworked **D**ata (uFIND). The purpose of uFIND is to provide rich and powerful facilities for the user to: discover and choose a desired dataset by navigation through the multi-dimensional metadata space using faceted search (Fig. 3.4); and seamlessly access and browse datasets with DataFed tools.

3.1.3 DataFed Workflow

The Service Oriented Architecture (SOA) is used to build web-applications by connecting the web service components (e.g. services for data access, transformation, fusion, rendering, etc.) in Lego-like assembly as illustrated in Figure 3.5 (Husar and Hoijarvi, 2008). The generic web-tools are also created in this manner include catalogs for finding, data browsers for spatial-temporal exploration, tools for transport analysis, spatial-temporal pattern analysis and download for use in other tools.

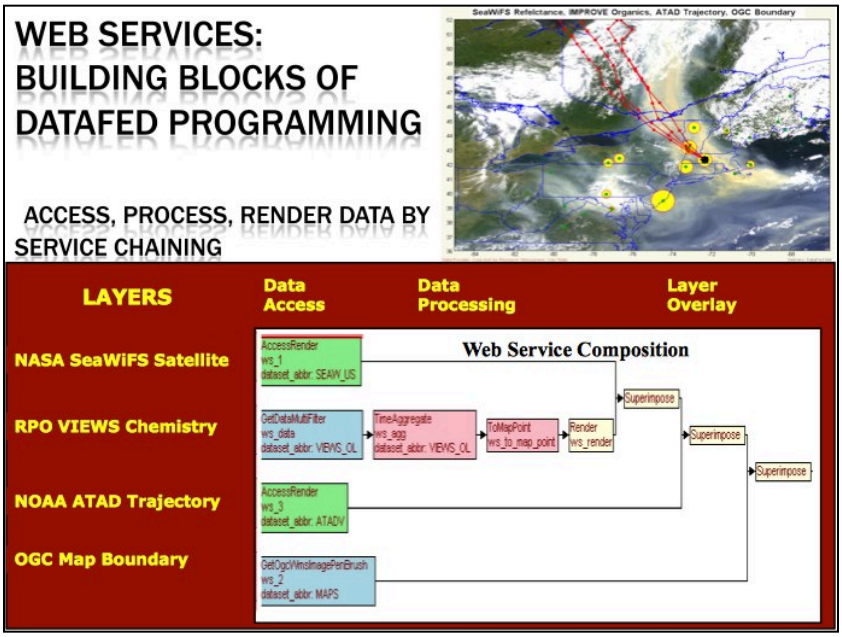


Figure 3.5 A typical web service program for the creation of multi-layer data views.

In DataFed, a custom-designed workflow engine using web service interfaces performs the orchestration of processing services. The workflow is designed for chaining both DataFed services as well as other, external web services. Likewise, DataFed's services are available to, and have been integrated with, other organization's workflow software. A data view (Fig. 3.5) is a user-specified representation of data accessible through DataFed. Data views consist of a stack of data layers, similar to the stack of spatial GIS data except that DataFed views can represent temporal and other dimensional pattern. Each data layer is created by chaining a set of web services, typically consisting of a Data Access Service, which is followed by the services for processing, portrayal etc. Data views are defined by an XML file, which contains the instructions to create a data view. The view file is also used to store the state, i.e. the input settings of the view.

The services are organized as a stack of workflow chains (Fig. 3.5). Each row is a data layer, where the values that are displayed are computed through the service chain. The service chain starts with data access followed by several processing services and then completed through a rendering service. Figure 3.5 illustrates a map view consisting of four independent data layers. The view in the top, right corner of Figure 7 shows the intrusion of forest fire smoke from Quebec to the N.E. United States. The color NASA satellite image is accessed

through an OGC WMS data access service. The point monitoring data are accessed from an SQL server through a wrapper, which formulates the SQL queries, based on the geographic bounding box, time range and parameter selection in the OGC WCS query.

The DataFed Browser/Editor, seen in Figure 3.6, is the primary tool for the exploration of spatial-temporal pattern of pollutants. The multi-dimensional data are sliced and displayed in spatial views (maps) and in temporal views (timeseries). Each data view also accepts user input for point and click navigation in the data space. Other views can also be displayed such as cyclic view for the display of diurnal, weekly and seasonal cycles at a given location or within a user-defined bounding box.

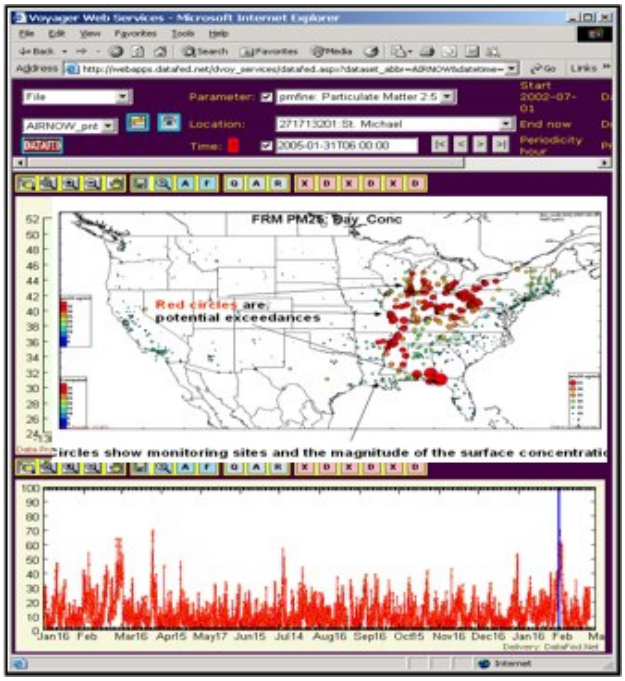


Figure 3.6 DataFed Browser for browsing data in space, time and parameter

3.2 Social Media as an Air Quality Sensor

As part of this thesis, the air quality information system was extended to gather air quality-relevant observations from social media. Social media sites like YouTube[v], Flickr[vi], Blogger[vii] and Delicious[viii] combined with Internet access, practically everywhere have lowered the threshold needed to share photos, videos, personal observations and other content in real-

time. These responsive human sensors with their smart phones have created a "skin" for the Earth.

Social media sites can be thought of as mediators like DataFed, that wrap distributed content into a web services and provide space to include metadata that aids in helping others find content. The social media sites allow for searching in multiple ways: keywords, text, who added the content, etc. One of the ways that these are "social" is by allowing the user to subscribe to the query via Really Simple Syndication (RSS). RSS is a format for allowing users to subscribe to receive updates whenever the content is changed. RSS has become a standard for publication through use and always includes: content title, description and link regardless of what site it comes from. This standardization allows RSSs from multiple sites to be aggregated or for tools like RSS readers to be built to read any feed. This combination of subscribing to a query via RSS allows for a traditionally, static query to become a real-time query that always provides the most current information.

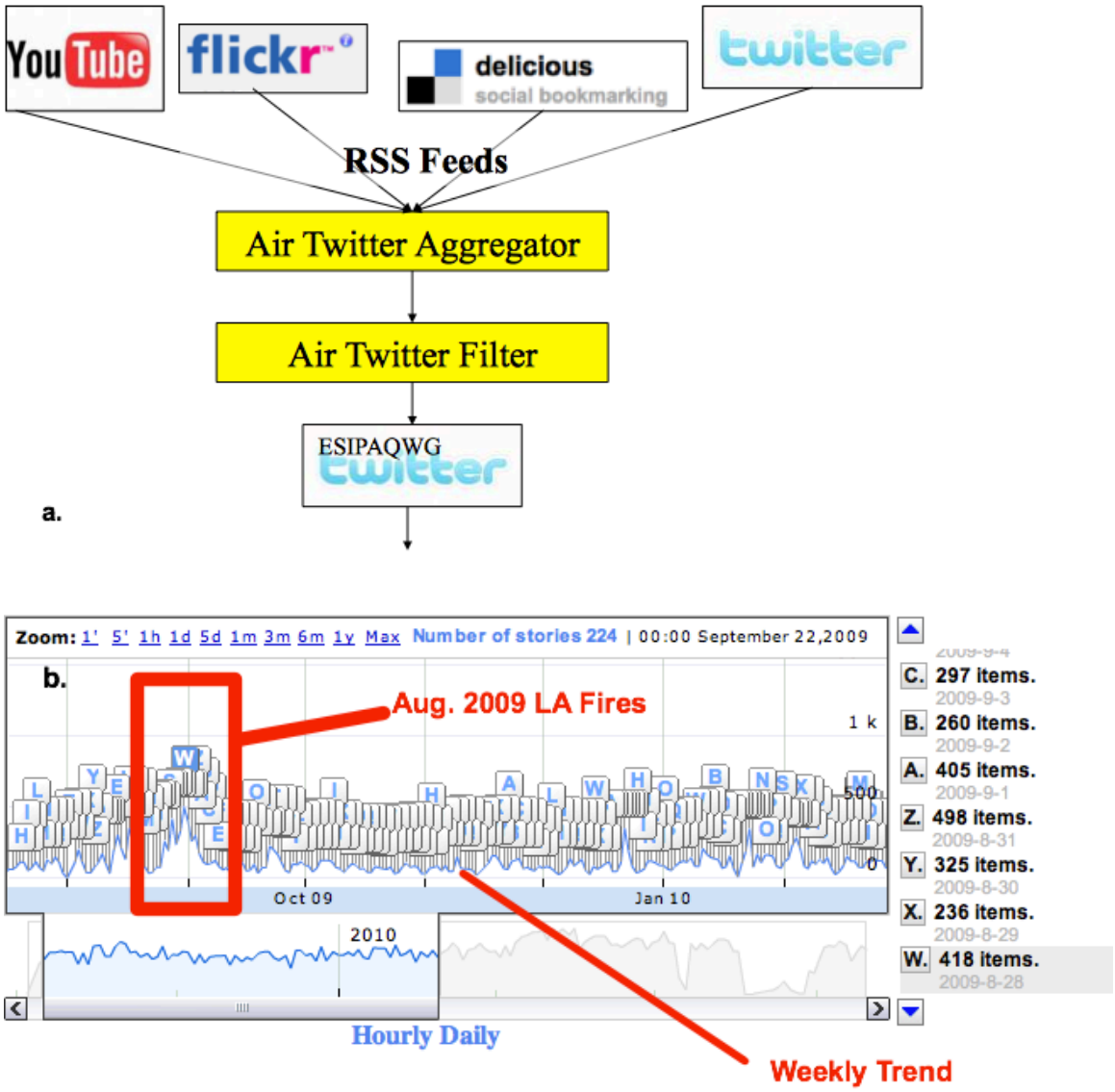
Businesses are taking advantage of aggregating RSS streams from various social media sites or "listening" to the social-media chatter about their brand, respond to complaints and develop brand loyalty by reaching out to customers. News agencies are also using social listening techniques and have implemented sites like iReport, since it is more and more likely that citizen reporters will 'break' news stories and identify major events.

Scientists can benefit from social listening as well. Air Quality (AQ) events such as fires and dust storms are highly visible and impact daily life, thus the pictures, videos, blogs and tweets are shared through web within minutes of the event occurring (Fig. 3.7). Figure 3.7 shows the results for a search on smoke from three major social media sources, Twitter, Flickr and YouTube during the August 2009 Southern California (SoCal) Fire. There were hundreds of videos and thousands of tweets describing the event.

Figure 3.7 YouTube, Flickr and Twitter displaying material published to the web during 2009 SoCal Fires

The air quality community is taking advantage of this new source of sensing information through, "community remote sensing". Community remote sensing incorporates the new and evolving social media 'sensors' along with remotely sensed surface and satellite data to provide contextual information about what is occurring in the environment.

Air Twitter (Robinson, 2010) is a social media listening tool that operates using an SOA approach (Fig. 3.8). AQ-related, user generated content is published on sites like Twitter, blogs, Delicious and Flickr and described using terms like air quality, fire and smoke. For the air quality application, we searched each of these social media sites for content tagged with 'smoke', 'dust', 'air quality' and other outdoor air quality-related terms. Once we found relevant search terms, we subscribed to the search list using the RSS link available from the social media site.



Tweets for 2009-8-28

- Gmorn yall! â
- So gross and smoky right now RT @lifeserial: Wildfire FAIL. Air quality is the suck.
- @geekgirldiva don't choke on any air quality on the way to the parking lot
- Headed for work. Going to try not to choke on the air quality on the way ;-)
- suppose to go swimming today. wondering how bad the air quality is, may have to nix it.
- @johncmayer How is the air quality there? In the SCV, I can't see the mountains through the smoke & it looks like it's snowing. Scary.
- soCal on fire :-O

Figure 3.8 a. Air Twitter Information System b. AirTwitter timeseries

The feeds from multiple social media sites are aggregated using existing aggregation services and then filtered to remove content that isn't relevant such as 'quality of Nike Air' for 'Air Quality', thus allowing Air Twitter to harvest outdoor air quality, user-generated content into a single stream. The Air Twitter stream is re-tweeted through a twitter account (@ESIPAQWG[ix]) (Fig. 3.8a). This feature of exposing the Air Twitter stream is a unique feature not common among business or news social listening. As a result of this exposure and constant sharing of real-time AQ information a community of over 1300 followers has developed. This community includes politicians as well as many local community AQ agencies that publish their real-time surface monitoring data through Twitter. Further, when AQ events do happen, this is the audience that we alert first.

The aggregated Air Twitter stream is also saved in a local database, which allows time series analysis[x] of the number of tweets hourly and daily in order to identify (Fig. 3.8b) AQ events. The letters above the time series are hyperlinks, clicking the letter will display the Air Twitter stream for the selected hour or day. As the number of tweets increase, we use this view to look for a common location or event thread in the tweets. The red box in Figure 3.8b highlights the increase in tweets seen in August 2009. This event identification occurs hours to a full day ahead of event identification only using scientific data.

3.3 Combining Science Data and Social Media for AQ Event Analysis

As AQ events are identified, collaborative EventSpaces (Robinson, 2008) are created to collect the social and scientific information about the event. An EventSpace[xi] is a collaborative, wiki workspaces, open to community editing and discussion. EventSpaces are devoted to a particular air quality event analysis to facilitate the community-based AQ analysis. These workspaces include both a structured and unstructured part. The structured part includes: when and where the event happened and what type of event occurred (smoke, dust, etc.). This allows for cataloging these events for future reference. The unstructured

part of the page is used to harvest community content such as pictures, video and blog posts about the event or AQ data from models, surface observations and satellites. Once the EventSpace is created, the @ESIPAQWG twitter account tweets the link to the relevant EventSpace wiki page every few hours in an effort to involve the broader AQ community that follows @ESIPAQWG.

Chapter 4

Case Study: Real-time Event Analysis of August 2009 Southern California Fires

Air Quality events are visible, dramatic environmental events. Both the social and scientific data cause AQ analysts to take note and since, at least at first, they are somewhat of a mystery there is significant willingness to share information and to work together. Demonstrations of this was seen in the Asian Dust event (Husar et. al. , 2001). Service-Oriented Air Quality analysis allows content from all over the web to be brought together as soon as the event is identified.

Analyzing the peak seen in the red box in Figure 10b, we saw that a majority of the tweets were about the Southern California Fire. This event identification occurred 24 hours ahead of satellite and surface monitors available for that area. This identification led to creating an EventSpace[xii] within the first day (Fig. 4.1) of significant burning. The link to the EventSpace was tweeted repeatedly through @ESIPAQWG to notify the AQ-interested followers and because of this publicity the link to the EventSpace was picked up by the LA Times and was also shared through California Fire-related groups on Flickr and CNN's iReport.

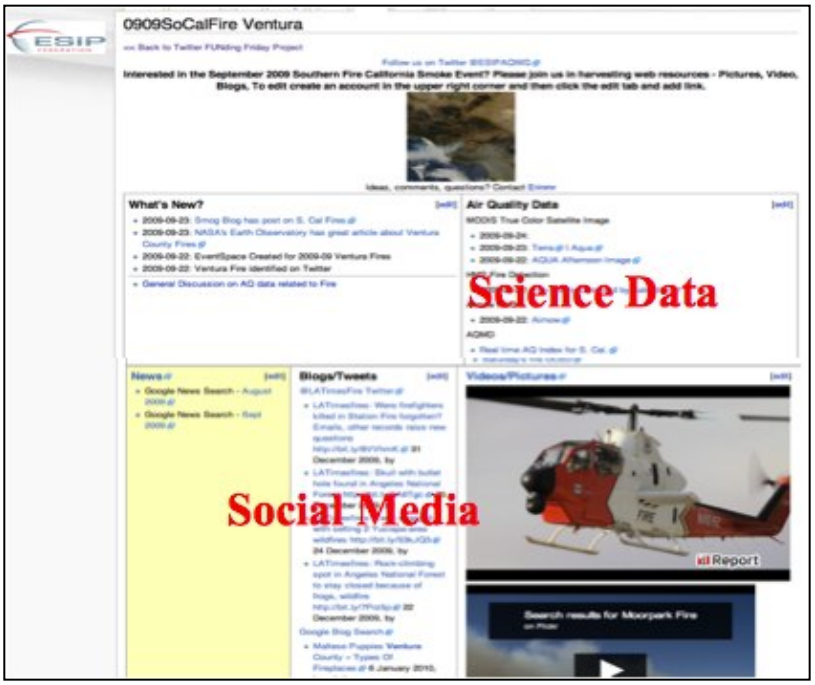


Figure 4.1 Air Quality EventSpace for 2009 Southern California Wildfires

The EventSpace was updated throughout the entire active period of burning with relevant content. As science data were available, days, weeks or months after the event, this EventSpace brought together relevant satellite data such as MODIS for true color images of the smoke, MODIS Fire pixels to identify fire locations and OMI Absorbing Aerosol Index to provide additional evidence for the spatial extent of the smoke. Surface observations from EPA's AIRNow were incorporated to identify the elevation of the smoke and models such as the Naval Research Laboratory's NAAPS smoke models provided further validation for the given spatial-temporal pattern of the smoke. All of the air quality data are accessed through the federated data system, DataFed (Husar, 2007) and displayed in the EventSpace using links to the DataFed views, screencast animations and kml layers for display on Google Earth.

The EventSpace was monitored using Google Analytics, which records web traffic and provides tools to visualize when and where the views are coming from. During the August California Fires the traffic increased five-fold to the ESIP wiki (Fig. 4.2a). Furthermore, the increase in traffic was entirely due to views of the SoCal Fire EventSpace (Fig. 4.2b). A top

driver to the site was from Twitter. Presumably, this was from the publicity gained by tweeting the link to the EventSpace to the @ESIPAQWG followers and having that link re-tweeted by major network nodes like the LA Times helped drive traffic up. An interesting and unexpected observation was that most of the increased traffic was coming from Southern California (Fig. 4.2c). So the right people were finding the right information at the right time.

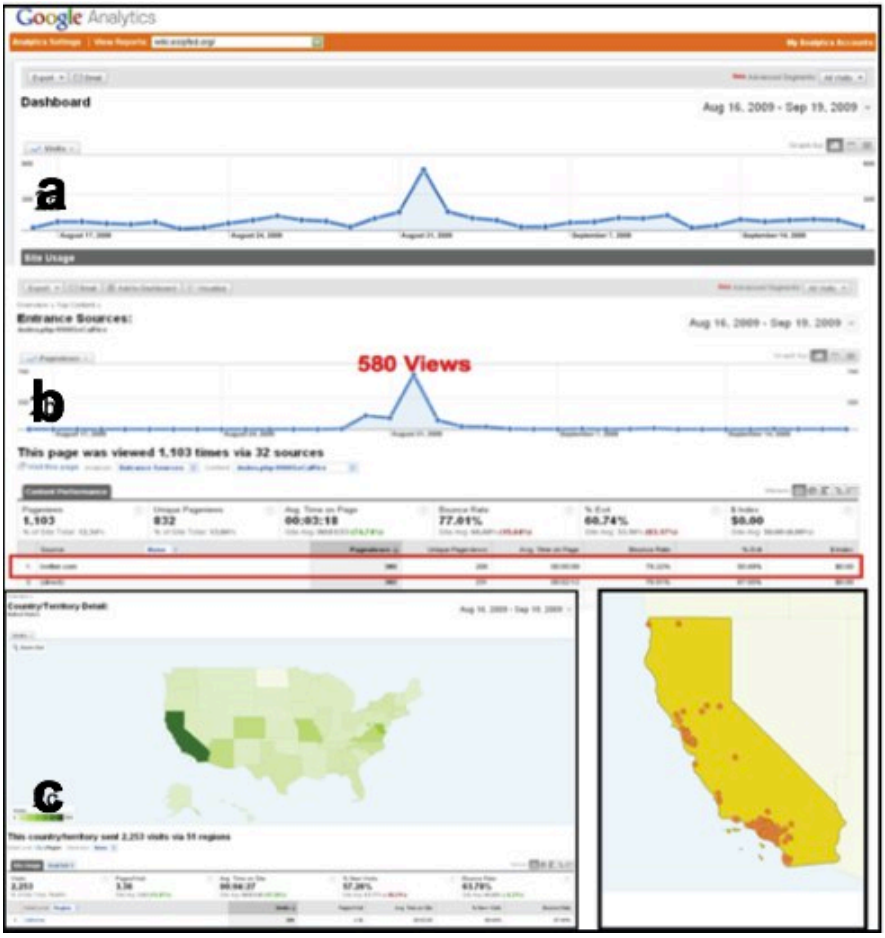


Figure 4.2 a. Entire site traffic for ESIP wiki Aug. 16-Sept.16; b. Traffic to SoCal EventSpace driven from Twitter; c. Geographic location of traffic.

The overall benefit of using the online community as an AQ event indicator, allows specific effort to be made for initial documentation of air quality events and the result is a catalog of descriptive observations with some sparse analysis that can be followed-up.

Chapter 5

Case Study: Exceptional Event Analysis for May 2007 Georgia Swamp Fires

The quality of ambient air is maintained at healthy levels by the setting and compliance with National Ambient air Quality Standards (NAAQS) based on measurements using Federal Reference Method (FRM) monitors. In the past, AQ regulatory decisions were made based on standard reference methods for $PM_{2.5}$, ozone, etc. In 2006, the NAAQS for $PM_{2.5}$ was significantly revised by reducing the daily standard from 65 to 35 $\mu g/m^3$ and recently for ozone from 85 to 75 ppb. The tightening of the short-term standards and the Exceptional Event (EE) Rule shifts the attention from controlling the yearly average to the reduction and control of short-term, episodic air pollution. Since the 2006 NAAQS amendments, both $PM_{2.5}$ and ozone are subject to the new EE Rule, which allows the exclusion of data strongly influenced by impacts from "exceptional events," such as smoke from a wildfire or dust from abnormally high winds. States "flag" data for those days that they believe to be impacted by exceptional events and must provide evidence for the event, using other data sources such as satellites and models. Such flagged days, if concurred with by EPA, may be given special consideration in the compliance calculations.

The flagging procedure has to be in accordance with section 40 CFR 50.14 (c)(3)(iii) of the EE rule. Preparing and evaluating the evidence for flagged data are a technically challenging task both for the State and the Regulatory offices. It requires:

A. Event Identification

B. Clear Causal Relationship between the Data and the Event

C. The Event is in Excess of the "Normal" Values

D. The Exceedance or Violation would not Occur, *But For* the Exceptional Event

The Exceptional Event Rule was the first EPA Rule where data from outside sources, such as media reports, satellite sensors and other surface observations as well as models could be incorporated to provide evidence. Fortunately, recent developments both in terms of service oriented data access and the new, collaborative web tools available offer outstanding opportunities to fulfill the information needs for the new agile air quality management approach.

5.1 May 2007 Georgia Swamp Fires

We will use the May 2007 Georgia Swamp Fire (Fig. 5.1) to illustrate this agile data system's ability to assess the impact of an exceptional event on the $PM_{2.5}$ concentration over the eastern U.S. using multiple datasets. The smoke from this event has impacted multi-state regions, which are also receptors of major anthropogenic sulfate sources. Hence it is a suitable event for demonstrating the key "but for" condition of the EE Rule (Husar and Robinson, 2008). The illustrative analysis will focus on May 24, 2007.

Figure 5.1 a. GA Smoke, b. Fire pixels c. Airnow surface $PM_{2.5}$ and OMI Absorbing Index

The images below are generated by web-based tools of DataFed: generic data exploration tools and three tools with specialized workflow were used for this analysis. All the tools leverage the benefits of OGC standards-based service oriented architecture: each tool is applicable to multiple datasets; service orchestration makes it easy to create new tools; and the shared web-based tools promote collaboration and communal data analysis. This allows the users to use these tools for the exploration of exceptional events that occur on other days.

5.1.1 A: Event Identification

The first step is to establish that a sample is a likely contributor to noncompliance. A site is in noncompliance if the 98th percentile of the $PM_{2.5}$ concentration over a three year period is over 35 $\mu g/m^3$. However, a sample may be in compliance even if the $PM_{2.5}$ concentration is > 35$\mu g/m^3$, provided that such values occur less than 2 percent of the time.

The $PM_{2.5}$ samples that are potential contributors to non-compliance can be determined visually and qualitatively by the $PM_{2.5}$ Data Browser Tool (Fig. 3.6). The map view shows the $PM_{2.5}$ concentration as colored circles for each station for a specific date. The coloring of the $PM_{2.5}$ concentration values (circles) is adjusted such that the concentrations above 35$\mu g/m^3$ are shown in red (Fig. 5.2). This provides an easy and obvious way to identify the candidate samples for noncompliance.

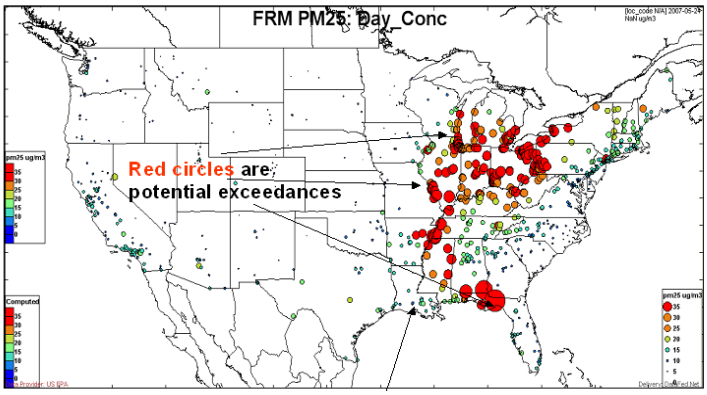


Figure 5.2 FRM data layer. Red circles show magnitude of surface concentration in exceedance

The second sub-activity of this step is to identify the exceptional event. An Analyst Console (Fig. 5.3) is a facility to display the state of the current aerosol system. These Analysts Consoles are key for establishing the emergence, evolution and dispersal of exceptional events. Through a collection of synchronized views data from a variety of disparate providers are brought together, the sampling time and spatial subset (zoom rectangle) for each dataset is synchronized, and that the user can customize the console's data content and format.

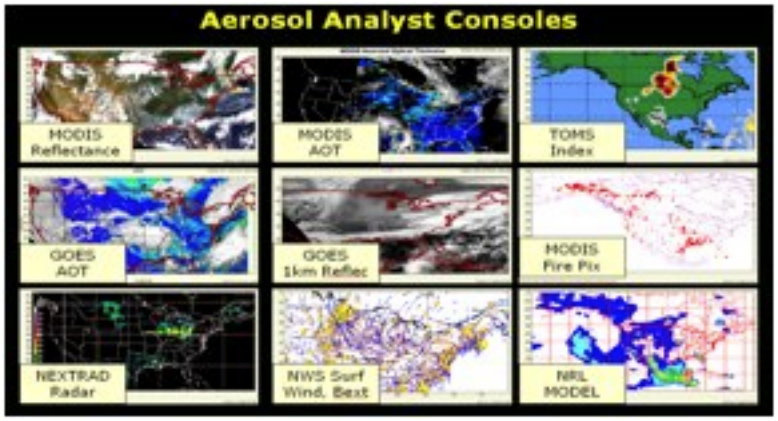


Figure 5.3 Analyst console showing a variety of data to provide spatial context for the event

Another way to identify and document events is through searching the web for social media and news articles. The general public provides additional qualitative observations of exceptional events shared through internet-accessible blog posts, photos through Flickr and videos through YouTube (Fig. 5.4). Figure 5.4 shows that at the time of analysis on the Georgia Smoke event, about a year later, there were hundreds of photos and videos and thousands of blog entries on the event.

Figure 5.4 Social Media available to document the event

In the future it is possible that these sources may be aggregated into EventSpaces described above in the real-time case study. This initial analysis could then provide a reference to state and regional analysts during this post-event analysis for event description.

5.1.2 B: Clear Causal Relationship between the Data and the Event

In the next steps, trajectory analysis is applied to delineate which of the monitoring sites are likely to be impacted by the smoke. In this analysis the location of the smoke source area is delineated by the black rectangle, centered on the Okefenokee fire location (Fig. 5.5).

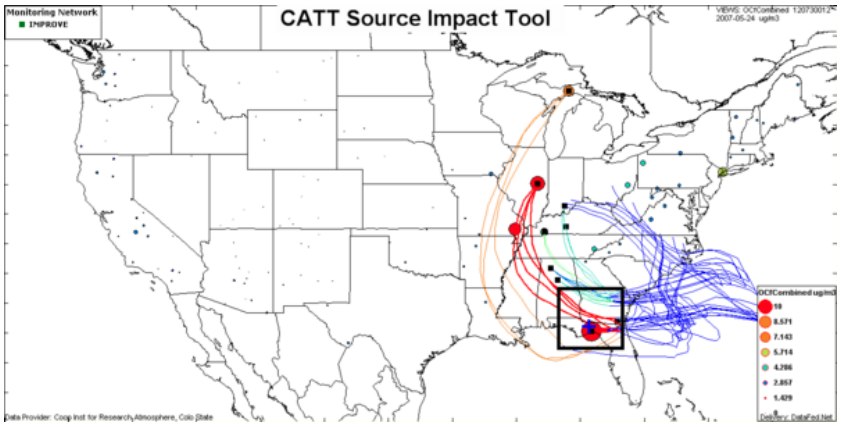


Figure 5.5 Backtrajectory analysis around Okefenokee Fire

All the backtrajectories that pass through that "source rectangle" are made visible while the other trajectories are suppressed. The coloration of individual trajectories prior to entering the source rectangle is set to thin blue lines. During and after the passage through the source rectangle, the trajectory line thickness and color is changed according to the concentration at the receptor site. By following the trajectories leaving the source rectangle, it is possible to delineate the regions of potential smoke impacts. That region can potentially be satisfying the "but for" condition. The backtrajectory analysis is done using the Combined Aerosol Trajectory Tool, CATT[xiii]. CATT is a web-based tool to explore the relationship between pollutant concentrations and their sources. It is based on ensemble backtrajectory aggregations for specific air-chemical conditions (Husar and Poirot, 2005).

Another sub-activity that is part of this step is to look at the chemical-spatial pattern of the event. The VIEWS chemical speciation data (Fig. 5.6a) and the NAAPS Model (Fig. 5.6c) from the Naval Research Laboratory both indicate that on May 24, the highest sulfate concentration was recorded just north of the Ohio River Valley. On the other hand, the

highest organic carbon concentration (Fig. 5.6b, d) is measured along the stretch from Georgia/Alabama to Wisconsin. This spatial separation of sulfate and organics indicates different source regions.

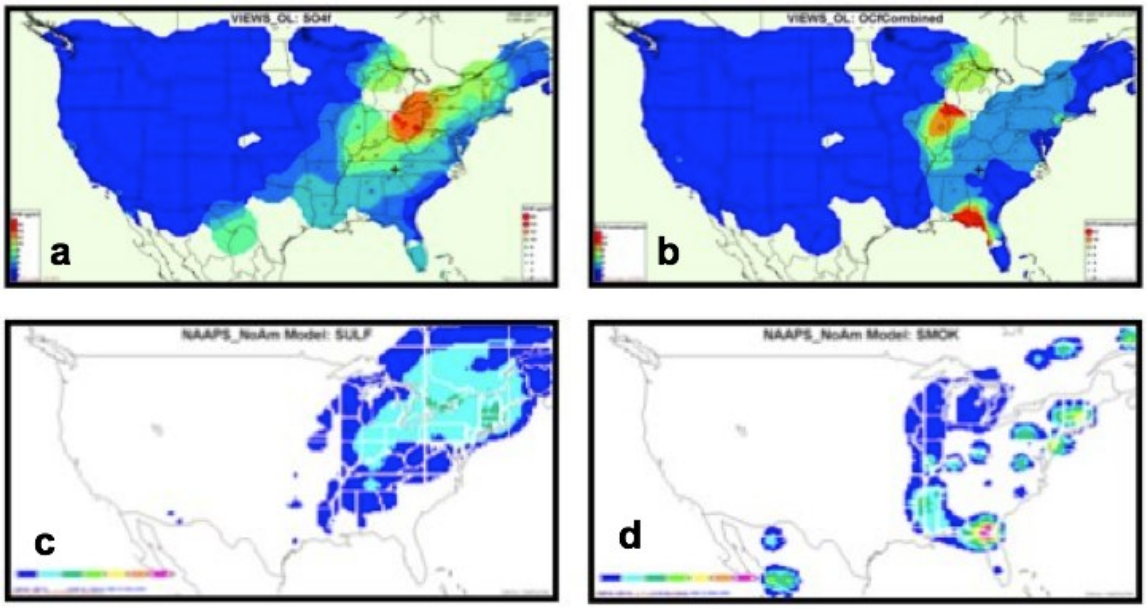


Figure 5.6 VIEWS a. Sulfate and b. Organic Carbon $PM_{2.5}$ speciation; NAAPS c. Sulfate and d. Smoke

5.1.3 C: The Event is in Excess of the "Normal" Values

A useful measure of the "normal" concentration is the 84th percentile (+1 sigma) for a given station. In the illustration below, a time windows of +/- 15 days (one month window) was chosen. This period is longer than a typical exceptional event, but it is sufficiently short to preserve seasonality. In order to establish the normal values the concentrations can be averaged over multiple years for the given time window measured in Julian days, i.e. days between 160 and 190 (Fig. 5.7).

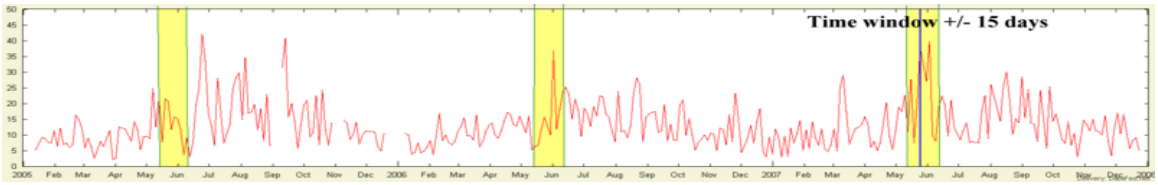


Figure 5.7 Three year time series for a site impacted by GA Smoke. Yellow indicates the +/- 15 day window.

Hence, a particular sample is considered anomolously high (deviates from the normal) if its value is substantially higher than the 84th percentile of the multi-year measurements for that "month" of the year.

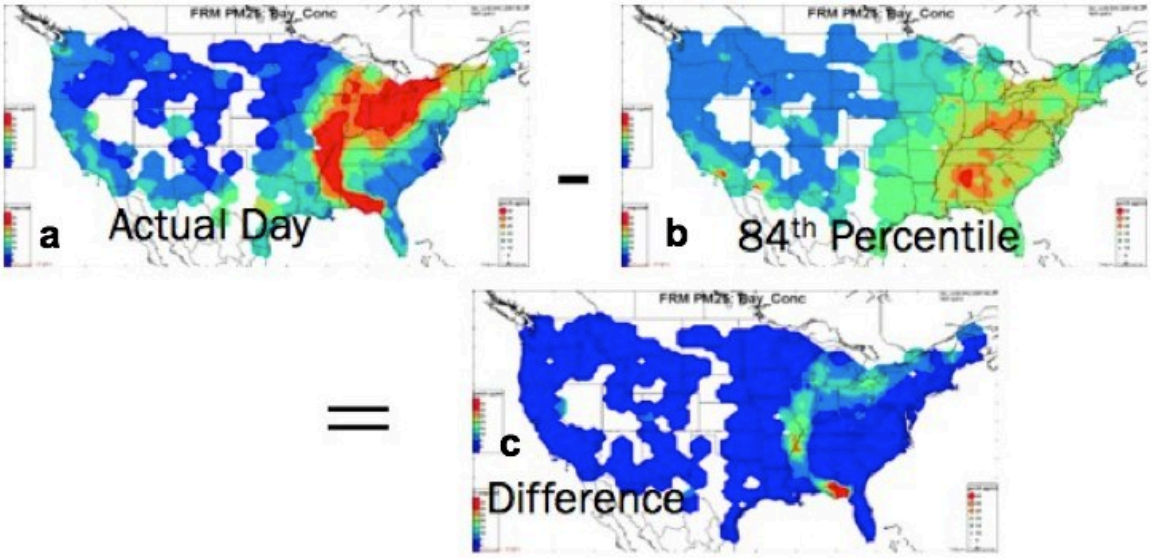


Figure 5.8 Anomaly detection a. actual measurement b. 84th percentile from 3 years c. difference.

Figure 5.8a shows the measured day average $PM_{2.5}$ concentration. Figure 5.8b shows the contour field for the 84th percentile $PM_{2.5}$ concentrations. Figure 5.8c shows the concentration anomaly, the excess concentration of the current day values over the 84th percentile values. Hence, a particular sample is considered anomalously high (deviates from the normal) if its value is substantially higher than the 84th percentile of the multi-year measurements for that "month" of the year.

This analysis was done using Concentration Anomaly Tool. This tool is a workflow chain in DataFed that provides a useful measure of the "normal" concentration. There is considerable need for flexibility in defining the 'normal' when calculating the deviation above normal and this tool allows defining the percentile used for normal, the number of days included in the analysis as well as modifying the dataset that the analysis is performed over.

5.1.4 D. The Exceedance or Violation would not Occur, But For the Exceptional Event

According to the EE Rule, observations can be EE-flagged if the violation is caused by the exceptional event. Considering the subtleties of the EE Rule, below are graphical illustrations (Fig. 5.9) of the Exceptional Event criteria.

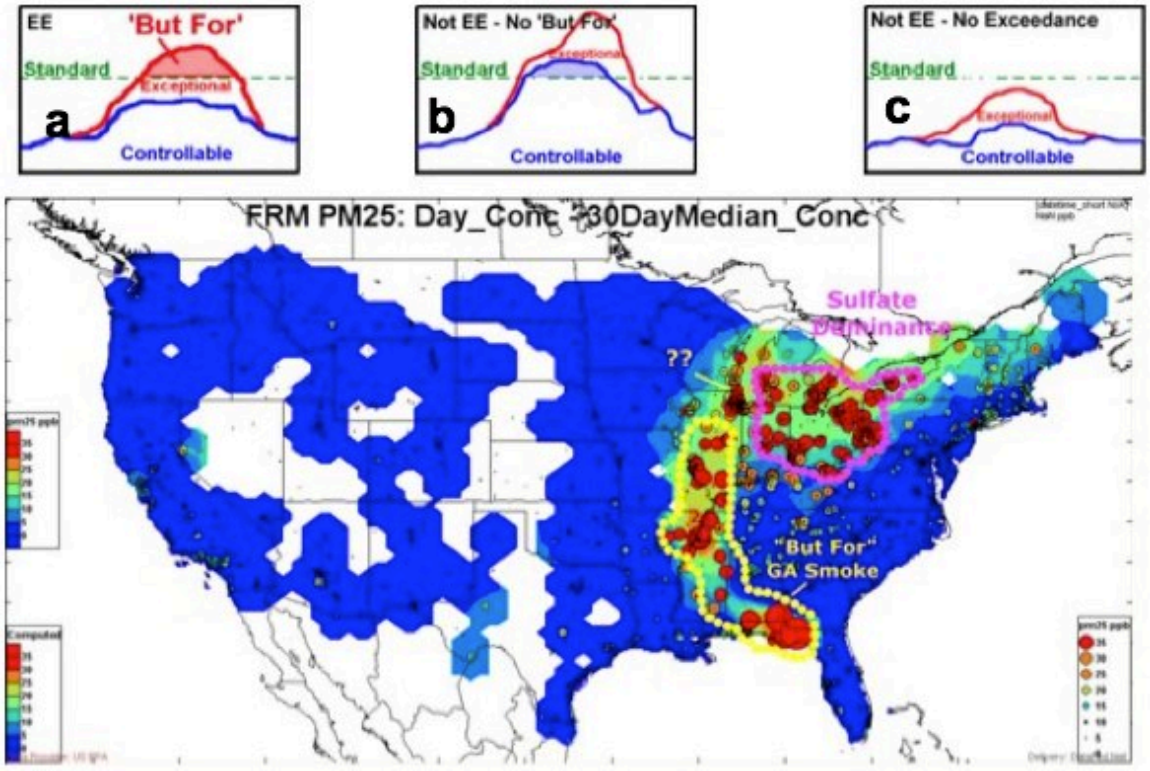


Figure 5.9 Areas identified for Exceptional Event status

- The leftmost figure (Fig. 5.9a) shows a case when the 'exceptional' concentration raises the level above the standard. A valid EE to be flagged.
- In the next case (Fig. 5.9b), the concentration from controllable sources is sufficient to cause the exceedance. This is not a 'but for' case and should not be flagged.
- In the third case (Fig. 5.9c), there is no exceedance. Hence, there is no justification for an EE flag.

Based on the combined chemical data and backtrajectories it is evident that the high PM concentrations that are observed along the western edge of the red trajectory path is due to the impact of the Georgia smoke. On the other hand the high $PM_{2.5}$ concentrations just north of the Ohio River Valley are primarily due to known, controllable sulfate sources.

5.1.5 Summary

For the May 24 Georgia Smoke event three different tools were used – the DataFed browser, CATT and the Concentration Anomaly Tool to manipulate the data services in order to provide the needed evidence. All of these tools were created using the service-oriented, workflow chains and can be reused and modified as needed. Air quality analysts did the bulk of the work, however, data were brought in from blogs and Flickr as evidence in the first step of event identification, so the public can be included as participants. For this analysis we used seven satellite datasets, three surface monitoring data, three weather datasets and one global model, totaling in fourteen datasets that were need. From this analysis (Fig. 5.9) it was possible to delineate the areas of the U.S. that were eligible for Exceptional Event status.

Chapter 6

Earth Observations Requirements

The section below is an initial attempt to quantify the data reuse through SOA by identifying the various Earth observation requirements. Earth Observation (EO) Requirements are the specific user needs for data and information needed for the AQ analysis. The process of EO Requirement identification (Husar, 2010) is:

1. Identify sub-activities needed to perform either real-time or post-event analysis
2. Identify the User Types needed for the sub-activity
3. Identify the EO's needed for each user

The two case studies show air quality analysis as the event happens in real-time and post-event as a regulatory procedure. However, both types of analysis need the same kind of information from social and scientific datasets. The difference is that in real-time analysis observations are staggered in availability. However, because of real-time event analysis a better effort could be made to acquire all of the necessary observations needed for post-analysis.

The following charts apply the above methodology to the Earth observation requirements needed for AQ Event Analysis. This analysis of EO needs allows identification of reuse for one type of measurement in multiple applications and where one application needs many observations. For air quality analysis 68 different requirements were identified[xiv]. Figure 6.1 shows that the requirement distribution among the four EE steps. It is clear that event identification and description and determining causality between a particular monitor and an exceptional event are the most EO intense. It is also clear that, except for the third step identifying the temporal anomaly, all of these analyses need more than just the FRM data.

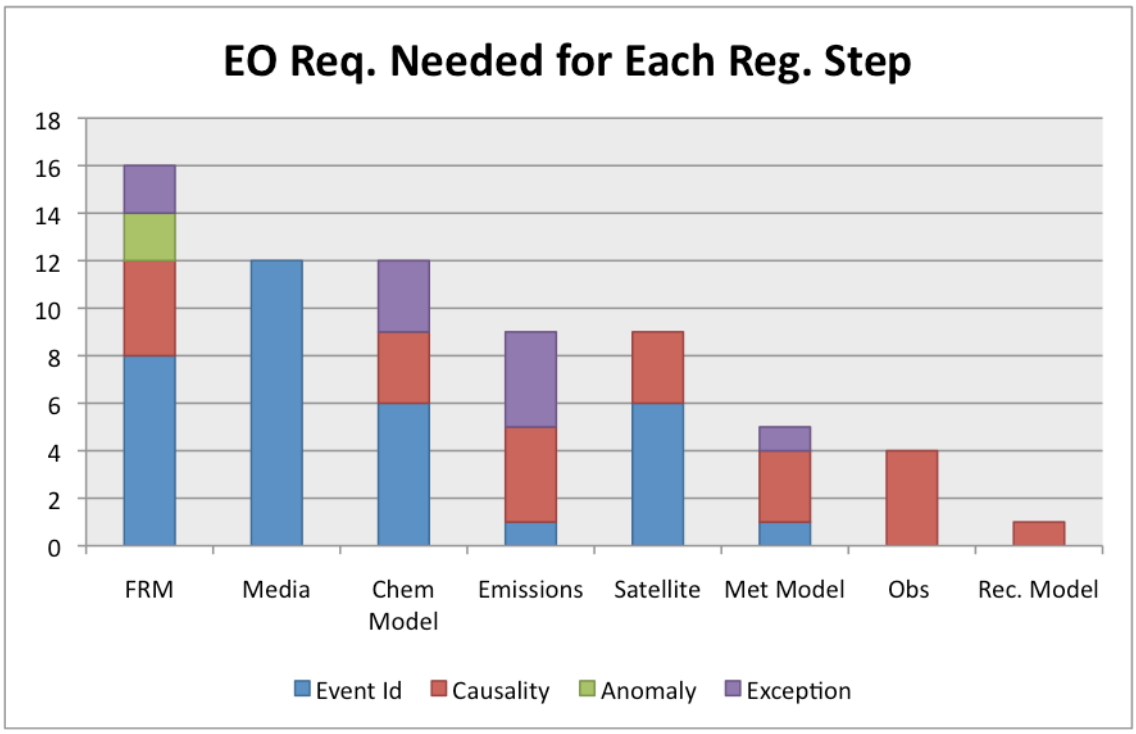


Figure 6.1 Earth Observation requirement by EE Rule Step

The above data shows the potential for data re-use. For example, FRM data can be reused four times, first identifying the event, establishing causality, then establishing that the event is an anomaly in time and finally that it is an exceptional.

Figure 6.2 shows the distribution of EO Requirements by user type.

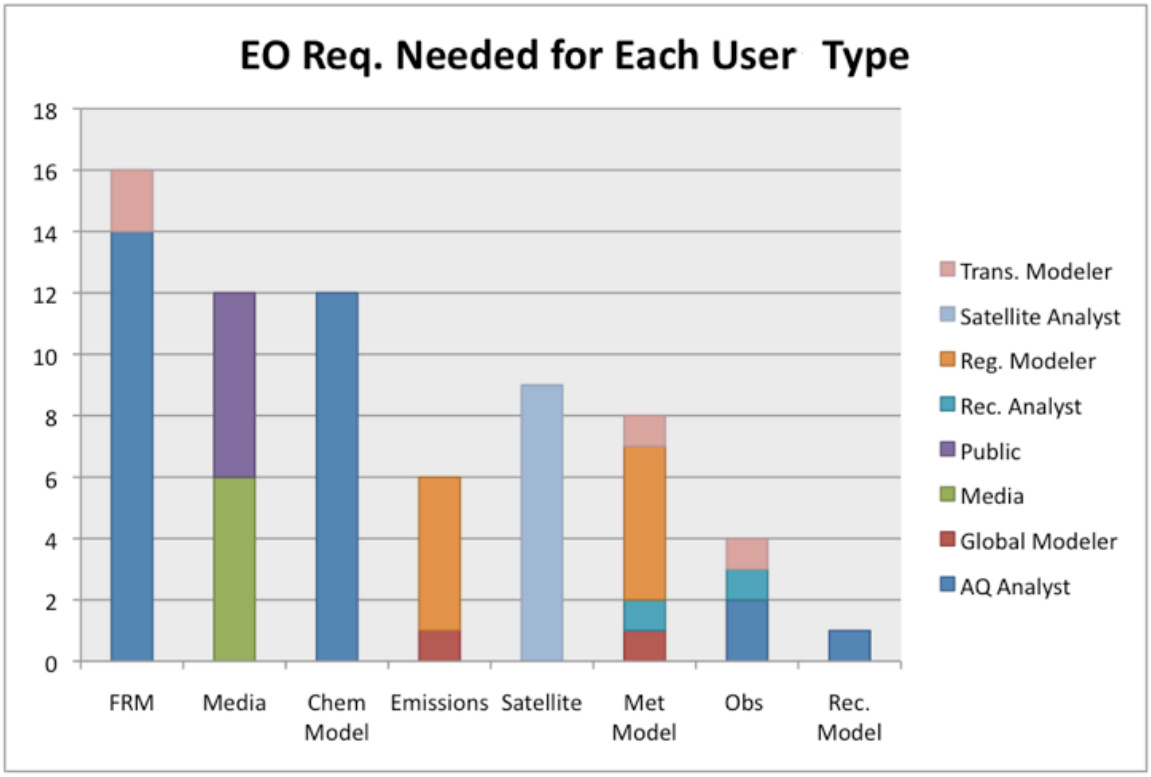


Figure 6.2 EO Requirement by User Type

This figure illustrates that once additional data are brought in, there are also numerous other types of users that are incorporated. These users may have no knowledge of being part of the Exceptional Event analysis however, because they are publishing their data (or event images) for their own purpose. The distribution of methods shows that the EOs needed come from a variety of sources.

Chapter 7

Future Work

This thesis has shown that AQ events need many different kinds of data for characterization and even with all the data the work is done best when it includes multiple perspectives. The future work includes improving collaboration within the AQ community to better share data and to move beyond data sharing issues to pursue more community-based event analysis. The other extension of this work is to expand beyond AQ. The SOA methodology scales up and down for any application that has distributed data that may be useful in multiple applications.

7.1 Collaboration on AQ Event Analysis

Full understanding and characterization of air pollution events is a very labor-intensive, subjective and sporadic process. Collecting and harmonizing the variety of data sources, describing events in a coherent, compatible manner and assuring that significant events will not 'fall through the cracks' is a challenging task for research groups, but even more for State and Regional air quality analysts. Conceivably, the event analysis performed in the community EventSpace could serve as triggers and guides to the States in deciding which station-data to flag.

7.2 Global Earth Observing System of Systems

Since 2005, a voluntary partnership has formed between 58 countries in order to share data in an effort to solve broad societal challenges. The Group on Earth Observations (GEO) is coordinating this partnership. GEO is an intergovernmental organization working to improve the availability, access, and use of EOs to benefit society. GEO is coordinating efforts to build a Global Earth Observation System of Systems (GEOSS, 2005). The only

way to create this loosely-coupled, dynamic System of Systems is to extend the SOA principles to a global level (Fig. 7.1).

Figure 7.1 Global Earth Observing System of System (GEOSS, 2005)

The main idea of GEOSS is that a single earth observation can have many uses and one societal benefit needs many observations. The GEO Information System, GEOSS, proposes to be a broker where data providers can publish their data and users can come to find data, access and apply the offered data in their respective social benefit areas. GEOSS is built using the same OGC standards used by DataFed, so once users find the data they will be able to directly access the data and use it in standards-based tools like DataFed.

References

Anthes, R. A, and A. Charo. 2005. *Earth science and applications from space: urgent needs and opportunities to serve the nation*. Natl Academy Pr.

GEOSS, GEO. 2005. *10-Years Implementation Plan Reference Document*. ESA Publications Division, Netherlands.

Husar, R., and R. Poirot. 2005. DataFed and FASTNET: Tools for agile air quality analysis. *EM-PITTSBURGH-AIR AND WASTE MANAGEMENT ASSOCIATION-*: 39.

Husar, R., D. Tratt, B. Schichtel, S. Falke, F. Li, D. Jaffe, S. Gasso, et al. 2001. Asian dust events of April 1998. *Journal of Geophysical research-atmospheres* 106, no. 16: 18317–18330.

Husar, R. B, and K. Hoijarvi. 2008. DataFed: Mediated web services for distributed air quality data access and processing. In *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*, 4016–4020.

Husar, R. B, K. Hoijarvi, and S. R Falke. 2008. DataFed: An Architecture for Federating Atmospheric Data for GEOSS. *IEEE Systems Journal* 2, no. 3: 366–373.

Husar, R. B, K. Hoijarvi, R. Poirot, S. Kayin, K. Gebhart, B. A. Schichtel, and W. Malm. n.d. *Combined Aerosol Trajectory Tool, CATT: Status Report on Tools Development and Use*. Paper.

Husar, R.B. 2010. URR AQ Management presented at the GEOSS User Requirement Registry Analyst Meeting, October 4, Washington, D.C.

Husar, R.B., and E.M. Robinson. 2008. Evidence for Flagging Exceptional Events. http://wiki.esipfed.org/index.php?title=Evidence_for_Flagging_Exceptional_Events.

McConnell, J.M. 2008. *Vision 2015: A Globally Networked and Integrated Intelligence Enterprise*. Defense Technical Information Center.

Peppler, R. A., C. P. Bahrmann, J. C Barnard, N. S. Laulainen, D. D. Turner, J. R. Campbell, D. L. Hlavka, et al. 2000. ARM southern Great Plains site observations of the smoke pall associated with the 1998 Central American fires. *Bulletin of the American Meteorological Society* 81, no. 11: 2563–2591.

Raffuse, S. M. 2003. Estimation of daily surface reflectance over the United States from the SEAWifS sensor. Washington University.

Robinson, E. M., K. Hoijarvi, S. Falke, E. Fialkowski, M. Kieffer, and R. B. Husar. 2008. Enabling Tools and Methods for International, Inter-disciplinary and Educational Collaboration. In *AGU Spring Meeting Abstracts*, 1:02.

Robinson, E. M. 2010. Air Twitter: Mashing Crowdsourced Air Quality Event Identification with Scientific Earth Observations. In *AGU Fall Meeting Abstracts*.

.

Vita

Erin Robinson

Date of Birth	September 10, 1984
Place of Birth	Raleigh, North Carolina
Degrees	B.S. Applied Science, May 2006 M.S. Energy, Environmental and Chemical Engineering, December 2010

December 2010

[i] http://datafedwiki.wustl.edu/index.php/VIEWS
[ii] http://www.opengeospatial.org/standards
[iii] http://datafed.net/
[iv] http://webapps.datafed.net/aq_ufind.aspx
[v] http://www.youtube.com/
[vi] http://www.flickr.com/
[vii] http://www.blogger.com/
[viii] http://www.delicious.com/
[ix] http://twitter.com/#!/esipaqwg
[x] http://pocus.wustl.edu/AQEar/graph.html
[xi] http://wiki.esipfed.org/index.php/ExceptionalEventList
[xii] http://wiki.esipfed.org/index.php/0908SoCalFire
[xiii] http://datafedwiki.wustl.edu/index.php/CATT
[xiv] http://webapps.datafed.net/AQ_needs.uFIND?table=http://datafedwiki.wustl.edu/images/3/32/EE_UR_102710.xls&title=Exceptional+Events+User+Requirements&usertype=aq%2banalyst