BioenergyKDF: Enabling Spatiotemporal Data Synthesis and Research Collaboration

Aaron Myers[1], Sunil Movva[2], Rajasekar Karthik[1], Budhendra Bhaduri[1], Devin White[1],
Neil Thomas[1], Adrian Chase[3]
1Computational Sciences & Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831
2 Cerner Corporation, 2800 Rockcreek Parkway, North Kansas City, MO 64117
3 School of Human Evolution and Social Change, Arizona State University, Phoenix, AZ 85004
myersat@ornl.gov, sunil.movva@cerner.com, karthikr@ornl.gov, bhaduribl@ornl.gov,
whiteda1@ornl.gov, thomasna@ornl.gov, adrian.chase@asu.edu

ABSTRACT

The Bioenergy Knowledge Discovery Framework (BioenergyKDF) is a scalable, web-based collaborative environment for scientists working on bioenergy related research in which the connections between data, literature, and models can be explored and more clearly understood. The fully-operational and deployed system, built on multiple open source libraries and architectures, stores contributions from the community of practice and makes them easy to find, but that is just its base functionality. The BioenergyKDF provides a national spatiotemporal decision support capability that enables data sharing, analysis, modeling, and visualization as well as fosters the development and management of the U.S. bioenergy infrastructure, which is an essential component of the national energy infrastructure. The BioenergyKDF is built on a flexible, customizable platform that can be extended to support the requirements of any user community—especially those that work with spatiotemporal data. While there are several community data-sharing software platforms available, some developed and distributed by national governments, none of them have the full suite of capabilities available in BioenergyKDF. For example, this component-based platform and database independent architecture allows it to be quickly deployed to existing infrastructure and to connect to existing data repositories (spatial or otherwise). As new data, analysis, and features are added; the BioenergyKDF will help lead research and support decisions concerning bioenergy into the future, but will also enable the development and growth of additional communities of practice both inside and outside of the Department of Energy. These communities will be able to leverage the substantial investment the agency has made in the KDF platform to quickly stand up systems that are customized to their data and research needs.

Categories and Subject Descriptors

H.3.5 [**Online Information Service**]: Data Sharing, Web-based Services; H.5.3 [**Group and Organization Interfaces**]: Web-based interaction

SIGSPATIAL '14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA
Copyright 2014 ACM 978-1-4503-3131-9/14/11…$15.00
http://dx.doi.org/10.1145/2666310.2666488

Keywords

Cyberinfrastructure, Knowledge Discovery, Spatial Data Infrastructure, Critical Infrastructure, Spatial Visualization, Open-Source, Web-Based, Communities Of Practice, Bioenergy.

1. INTRODUCTION

In many fields of scientific research a separation exists between publications and the data, models, and tools used in research. Typical ways to address this include digital assets registries and scientific data repositories; however, the purpose of these systems is generally to support metadata searching and data storage. Additionally, many instances in which scientists write about their analysis do not result in peer-reviewed publications (e.g. sponsored deliverables to DOE), and most datasets referenced in publications are not registered in these large repositories. This creates systematic issues where research can be funded, conducted, and then lost - sitting on a hard drive on the researcher's desk where it cannot be accessed and shared with the community at large. The BioenergyKDF provides digital data curation and data sharing among a community of researchers in order to prevent these issues by bring publications, data, models, and tools together under a single platform.

The Bioenergy Knowledge Discovery Framework (BioenergyKDF), found at https://bioenergykdf.net, provides national spatiotemporal decision support capability through comprehensive data and modeling resources that support analysis, visualization, sharing of data for researchers and a variety of stakeholders across public and private sectors for efficient planning, development, and management of the U.S. bioenergy infrastructure. This framework allows users to access tools through a web-based interface that support analysis, synthesis, and visualization of data that facilitates informed decision making by Bioenergy Technologies Office and a variety of public and industry stakeholders. The intention of the BioenergyKDF is to further research by sharing current data, models, and publications as well as archive results from previous studies as well as highlighting the strategic and tactical issues that need to be addressed with data and models appropriate and applicable at the spatial and temporal scales appropriate to those the issues. Moreover, this system is designed to be functionally and technologically sustainable for continued support of the Nation's Bioenergy mission for the coming decades.

2. RELATED WORK

Other web platforms have successfully provided similar features for different research communities. Data.gov "increases access to high value, machine readable datasets generate by the Executive Branch of the Federal Government" for various research communities including climate, agriculture, health, and energy [1]. This presidential initiative opens data access for difficult to find information, data, and services. One shortcoming is a lack of integrated exploration and analysis tools to further understanding of the data of interest.

HUBzero, OpenABM, and tDAR are other prominent platforms that powers Communities of Practice (CoP) sites [2, 3, 4]. While some of these platforms are now open-sourced and share their foundational frameworks, BioenergyKDF was built before the code was made available (i.e. HUBzero) or the community features were fully functional (i.e. OpenABM). As such, BioKDF required independent technology integration.

3. SYSTEM ARCHITECTURE

The BioenergyKDF is built on a large collection of open source technologies primarily Drupal, PostgreSQL with PostGIS, Apache SOLR, GeoServer, OpenLayers and ExtJS. An overview of the system architecture is illustrated in Figure 1.

3.1 Content Management System

A heavily customized Drupal instance backed by PostGIS and PostgreSQL database is the core of BioenergyKDF. The core manages all the content (data and meta-data) along with user identity/access management and web interface. The core also provides SOAP and REST based web services for serving the different components of the web interface.

Drupal is a highly modular, scalable and feature-rich open source content management system (CMS). It powers over a million sites including many Government and Top 500 Fortune company websites [5, 6]. Drupal offers wide variety of features expected from a modern CMS. It provides a large collection of APIs to extend its core, such as search, content types, menus, caching, logging etc. Separation of design, content and structure makes appearance and structure modifications easier without the need to touch the content [5, 6]. Beyond this core, Drupal's community has a huge base of contributed modules. Security is another benefit, as its codebase rigorously and constantly tested. Drupal's emphasis on open and well-established standards, makes it equipped to handle various needs such as data retrieval and consumption, and search engine optimization (SEO) [5, 6].

3.2 Data Services

Apache Solr is a scalable enterprise search platform due to its extremely fast and feature-rich search capabilities. It's built on top of Apache Lucene, one of the widely recognized and used indexing libraries. Major features of Solr include full-text search, faceted search and filtering, rich document handling and dynamic clustering. It provides an extensible plugin-based architecture, making it attractive for adoption [7, 8].

Solr support Geospatial searches out of the box – (1) Index points, shapes, and numeric duration such as multi-value time, (2) Boost search result scores based on distance, and (3) Results can be filtered by shapes such as bounding box [8]. BioenergyKDF periodically indexes all the content it manages in an Apache Solr instance that resides alongside the CMS. It builds on top of Apachesolr module in Drupal to facilitate this management [7]. This index is the backbone of the library section of the web interface. Also, the BioenergyKDF web-mapping server is built on **GeoServer** and sharing the common PostgreSQL database.

3.3 Map Interface

The map component of the web interface is based on OpenLayers and the ExtJS JavaScript framework. All requests from the map component are sent to the CMS where it proxies them to the BioenergyKDF web mapping server upon evaluating the logged in users access permissions to the requested resource.

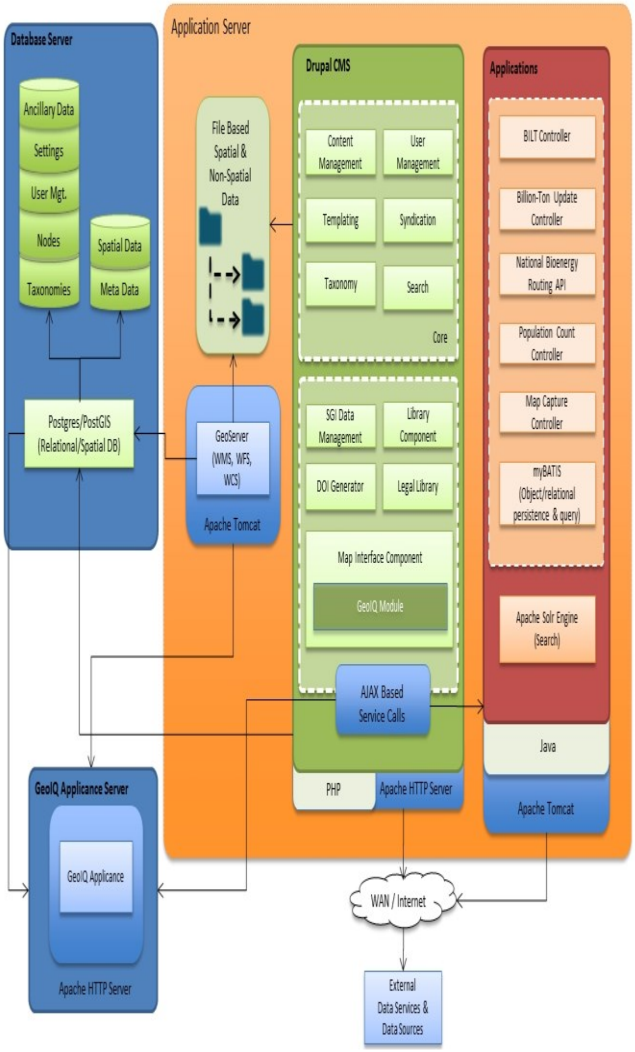

Figure 1: System Architecture of BioenergyKDF

OpenLayers 2 is a popular open-source and client side JavaScript libraries for making interactive web maps viewable in any browser. It provides various utilities needed to create mapping applications with the ability to easily customize many aspects of a map. It supports various map server back ends, notably GeoServer, ArcGIS Server, and MapServer [9].

ExtJS 4 is one of the leading JavaScript application frameworks for developing rich and interactive web applications by using Model-View-Controller (MVC) architecture. It provides a highly optimized class system, data package, layout managers and multitude of plugins such as grid and tree views, forms, drawing, charting, widgets etc. to create rich web applications. Dynamic Loading, drag and drop, localization, unit testing, cross-browser compatibility, debugging tools, and good documentation are other key benefits [10]. Custom **Drupal modules** have been developed for tighter integration, and look and feel with our mapping application. Sencha CMD, Drush and Linux shell scripts, Apache Tomcat are some other tools used for lifecycle management of our system.

4. SYSTEM CAPABILTIES

4.1 Research Collaboration

The BioenergyKDF supports role-based user access to a variety of local and externally served data sources. Users can quickly discover data holdings using both a single field, Google-style search or by filtering based on predefined categories. Users also have the ability to upload and manage data for public use or within groups where data can be shared privately. By providing a venue for users to advertise their own data and services, and also providing data provision and visualization capabilities for users who have not developed this resource on their own - the BioenergyKDF provides a rich environment for collaboration and research within the bioenergy community.

One unique capability of the BioenergyKDF is the capability to associate various data items with each other in order to create data or knowledge "packages". The highest level of categorization involves separating data items into one of three information categories: data, model, or literature. Information from the same research effort, or research of a similar nature, can then be organized into these information categories. By associating data, the user can quickly draw connections among data generated from various research efforts. By instigating a search, for example, with the model used to create a dataset the user can then follow those results to other publications containing similar data or research. Traditionally, getting access to data references in a publication has been challenging because journals do not require the raw data for publication and there are few venues to easily post data on the web; associating data items fills this gap.

4.2 Geospatial Analysis and Visualization

In addition to data management and search capabilities, the BioenergyKDF also offers a robust geospatial data viewer that includes support for spatial analysis and querying capabilities. The purpose of this capability is to allow the investigation of data interactions in geographic space using data pulled from disparate data sources within a single interface.

The BioenergyKDF can query this information in a variety of ways depending on the source data service type. A spatial identify feature, which will query based upon a single point selection on the map is possible on all datasets. A more robust bounding box query is also available that allows the user to query the data by drawing a box on the map. All features that intersect the defined area are returned and available for viewing within a sortable table that can be exported to an Excel Spreadsheet. Additionally, the user can query datasets via their attributes using a SQL like query builder.

Moving beyond simply querying datasets, the BioenergyKDF is also flexible enough to interact with spatial data models and analysis. An example of this type has been created for demonstration purposes and allows the user to interactively use a multi-modal Bioenergy Routing Model. This functionality helps to bring new analytic capabilities from the desktop to the web and to allow non-expert users to benefit from these models and to include the results in their own research. Another example is presented in the form of a tool that calculates the population within the displayed map area using the LandScan Global population dataset [11].

4.3 Research Models

Models such as "Billion-Ton Update Data Explorer and Download", "National Bioenergy Routing" etc. have been integrated into our map so users can quickly run, visualize and download data. Model results run by the user are saved in BioenergyKDF servers for future uses.

One of the corner stone capabilities that are helping to fulfill the mission of the KDF is the dissemination and exploration of the 2012 **Billion-Ton Update (BT2)** Study [12]. The BT2 reports builds off of "*Biomass as a Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply"* by providing updates to "a county-by-county inventory of primary feedstocks, price and availability quantities for individual feedstocks, and a more rigorous treatment and modeling of resource sustainability" [12, 13]. In order to assist with the promotion, dissemination, and data exploration it was decided to create interfaces within the KDF that would allow users access to the full data downloads, exploration of county and state level maps, and the export of custom data downloads.

The BT2 data consists of 59,264,818 data records split across 3 categories for Primary Agriculture (46,441,548 records), Forest Resources (11,050,780 records), and Secondary Resources (1,772,490 records), which includes aggregation to state level, but not conversion to metric units that occurs via database views. Managing and organizing this data to be easily queried and rapidly accessed for map display was one of the challenges in building this capability.

The BT2 Data Explorer provides users with a means to explore the BT2 data for varying scenarios and returns map layers for the provided parameters. One such way uses "wizard" like workflow to select resources of interest, scenarios, prices, and feedstocks. Results of the dataset will be displayed on the map to allow for quick visual inspection of county or state level variation. The "Billion-Ton Update Download" is another way and easy-to-use tool that allows the user to download 2011 BT2 data based on their parameters. A spreadsheet with generated results is available for download instantly, with one sheet for each desired resource along with a summary page of the download criteria.

A unique benefit of the services-based architecture supporting the Bioenergy KDF is support for an "out of the box" Application Programming Interface (API) used to support data exchange. With the Billion-Ton Update data, this API was refined and made available for public consumption. This capability was announced at the 2012 White House Energy Datapalooza [14].

The bioenergy legislative library permits users to interactively search through multiple facets of bioenergy legislation in Congress. The library includes the legislators and committees that collectively proposed and authored that legislation. Each of these three types of information – legislation, legislators, and committees – is fully integrated and visible in the library.

Potentially interested policy makers can be identified and their legislative histories checked. The effectiveness of committees to propose and pass legislation can also be reviewed. In essence, the legislative library offers its users the ability to quickly find information on the policies and policy makers in Congress and with this information in hand tailor future funding and research efforts in bioenergy to better support policy makers and the legislation they create.

5. RESULTS

The BioenergyKDF was officially released in January 2011. There are currently 1207 registered users and the BioenergyKDF has received 55,200 unique site visits as of June 2014.

5.1 Return on Investment

One of the remaining research areas regarding the BioenergyKDF and other systems of this type is the return on investment (ROI) for investing in a collaborative, open-source framework. For the Billion-Ton Update tools, an ROI analysis was done to further understand the benefit of the BioengyKDF as it relates to data access and exploration. Table 1 provides a simple example of a quantifiable ROI for Billion-Ton Update, where a known dollar amount was being spent supporting a capability and moving that capability to the user reduced cost by an estimated value. However, further research is needed to better quantify the ROI of the system as a whole. For example, what is the value of a single page view? How much ROI for every minute a user spends on the site? This is an area where the BioenergyKDF team will focus its efforts to attempt to address some of these questions.

Table 1: ROI Calculations for Billion-Ton Update Data Access through the BioenergyKDF

Task	Time	User Needs	Savings @ $200 per hr.
Map Generation	1 hour / map	2,750 maps	$550,000
Predefined Maps	15 min / map	2,260 maps	$113,000
Data Access	15 mint / request	3,400 requests	$410,000
Total			$1,730,000

6. CONCLUSION

The BioenergyKDF is built upon a flexible, customizable platform that can be extended to support the requirements of any user community. Its component-based platform and database independent architecture allow it to be deployed to existing infrastructure and to connect to existing data repositories. As new data, analysis, and features are added; the BioenergyKDF will help lead research and support decisions concerning bioenergy into the future, but will also enable the development and growth of additional communities of practice both inside and outside of DOE. These communities will be able to leverage the substantial investment the agency has made in the KDF platform to quickly stand up systems that are customized to their data and research needs. This is already happening in areas as diverse as nuclear fuel storage and freight transportation modeling, whose research communities recently stood up instances of the KDF.

There are also larger efforts underway to create thematic umbrella platforms built on the same technology, under which will be located specific community sites that are linked together architecturally, thus enabling cross-community knowledge discovery, data synthesis, and research collaboration. Facilitating this kind of interaction amongst researchers at multiple levels is the ultimate goal of the KDF project.

In the future, the BioenergyKDF will continue to add new features and functionality building off social networking principles, improved spatial and non-spatial data visualization, advanced spatial analysis and modeling, and enhanced search. Next version of our mapping application is in early works that supports both web and mobile clients, using Node.js, ExtJS 5, OpenLayers 3, and HTML5 [15]. The BioenergyKDF will also provide automated knowledge discovery algorithms that will allow publications and documents to be associated with other similar information in an automated process upon ingestion into the system.

7. ACKNOWLEDGEMENTS

This manuscript has been authored by employees of UT- Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

8. REFERENCES

[1] Data.gov. URL http://www.data.gov

[2] M. McLennan, R. Kennell, "HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering," Computing in Science and Engineering, 12(2), pp. 48-52, March/April, 2010.

[3] Janssen, Marco A., et al. "Towards a community framework for agent-based modelling." *Journal of Artificial Societies and Social Simulation* 11.2 (2008): 6.

[4] McManamon, Francis, Keith Kintigh. 2010 Digital Antiquity: Transforming Archaeological Data Into Knowledge. The SAA Archaeological Record 23(2):37-40.

[5] VanDyk, J. 2008. Pro Drupal Development (2nd edition). Apress, New York.

[6] What is Drupal? URL http://www.acquia.com/what-is-drupal

[7] Apache Solr. URL http://lucene.apache.org/solr/

[8] Spatial Search. URL http://cwiki.apache.org/confluence/display/solr/

[9] Perez, A.S., 2012. OpenLayers Cookbook. Packt Publishing.

[10] Sencha. URL http://www.sencha.com.

[11] Bhaduri, B., Bright, E., Coleman, P., and Dobson, J. 2002. LandScan: locating people is what matters. Geoinformatics 5(2):34-37.

[12] U.S. Department of Energy. 2011. U.S. Billion-Ton Update: Biomass Supply for a Bioenergy and Bioproducts Industry. R.D. Perlack and B.J. Stokes (Leads), ORNL/TM-2011/224. Oak Ridge National Laboratory, Oak Ridge, TN. 227p.

[13] Perlak, R., Wright, L., Turhollow, A., Graham, R., Stokes, B., and Erbach, D. 2005. Biomass as a Feedstock for a Bioenergy and Bioprocessing Industry: The Technical Feasibility of a Billion-Ton Annual Supply, in Joint Study sponsored by USDOE and USDA., Oak Ridge National Laboratory.

[14] Energy Datapalooza Fact Sheet: Unleashing the Power of Open Data to Advance our Energy Future. URL http://www.whitehouse.gov/sites/default/files/microsites/ostp/energy_datapalooza_fact_sheet.pdf

[15] Karthik, R. and Lu, W. 2014. *Scaling an Urban Emergency Evacuation Framework: Challenges and Practices*. Big Data and Urban Informatics (BDUIC) 2014.