Davis, G., G. Madey, and the VectorBase Consortium, “Cyberinfrastructure: The Key to Building Successful Science Gateways”, *Science Gateway Institute Workshop - IEEE Cluster 2013*, Indianapolis, Indiana, September 27, 2013.

Cyberinfrastructure: The Key to Building Successful Science Gateways

VectorBase: A Bioinformatics Resource Center for Infectious Diseases

Gregory J. Davis
Center for Research Computing
University of Notre Dame
Notre Dame, IN, USA

Gregory R. Madey
Computer Science & Engineering
University of Notre Dame
Notre Dame, IN, USA

The VectorBase Consortium
EMBL - EBI
Harvard University
IMBB - Crete
Imperial College
University of New Mexico
University of Notre Dame

***Abstract*—The VectorBase Bioinformatics Resource Center (BRC) is a science gateway, funded by the National Institute of Allergies and Infectious Diseases (NIAID), entering its 10 years of service to the scientific community on invertebrate vectors that transmit human diseases. This abstract describes three key factors contributing to the success of this science gateway: virtual organization, data and services integration, and community involvement and outreach. The role each plays in how they are impacted by the development of cyberinfrastructure is discussed.**

Keywords—Science Gateway; Bioinformatics Resource Center; Infectious Diseases; NIAID; Arthropod Vectors; Cyberinfrastrucure

I. *INTRODUCTION*

The VectorBase Bioinformatics Resource Center (BRC) is a web-based science gateway dedicated to providing data and bioinformatics resources to the scientific community interested in invertebrate vectors that transmit human pathogens [2][3][4]. The BRC provides a forum for the discussion and distribution of news and information relevant to invertebrate vectors, as well as access to tools to facilitate the querying and analysis of the data sets presented by the BRC. The BRC is comprised of scientists, bioinformaticians, and scienfic programers who curate data, develop and maintain the resource, and provide outreach and support to the user community which consists of almost 1,000 research scientists (those subscribed to a news mailing list) and a broader group of interested users (over 50,000 unique visitors per 6 month period). The VectorBase BRC is entering its 10th serving these communities, and is anticipating support from the National Institute of Allergies and Infectious Diseases (NIAID) for another 5 year contract.

The success of such a science gateway can be measured in various ways: continued funding, increasing user base, publication citations, and so on. Regardless of the measure, complex gateways will succeed or fail depending on how well they serve their target community given finite resources. Below we discuss three key factors that have enabled the BRC to become a popular and well-regarded scientific gateway: virtual organization, data and services integration, and community involvement and outreach. Within the discussion of each, we describe how the use of cyberinfrastructure has facilitated and strengthened the BRC’s position on these factors.

II. VECTORBASE - A VIRTUAL ORGZNIZATION

The VectorBase BRC is a virtual organization (VO). A distributed team of over 25 persons, both part-time and full-time, from six international instutotions and is responsible for cyberinfrastructure development, data curation, generation of derived data, end-user outreach, management, training and support. Fig. 1 below shows the responsibilities of staff from the participating organizations.

Structuring the BRC as a VO potentially provides some advantages over a traditional organization. First, it allows the BRC to assemble the necessary expertise required to build and maintain a sufficiently complex scientific gateway regardless of geographic location or institutional affiliation. A critical mass of talent need not be confined to a particular region or institution. Additionally, the organization can flexibly adapt to accommodate the changing needs of the gateway’s community. These advantages translate into a more efficent use of limited resources.

Despite these advantages, there are challenges that accompany the VO structure. The primary challenge of a distributed organization is effective and timely communication. Many BRC team members participate on a part-time basis and are located in a variety of different time zones. This challenge is exacerbated by the fact that the BRC is committed to relatively rapid release schedule (bimonthly). Another challenge, though not unique to the VO structure, is finding mechanisms by which goals can be formalized, responsibilities assigned, progress tracked. These challenges, even in a

The VectorBase science gateway is supported by NIH/NIAID contract No. HHSN272200900039C to the University of Notre Dame.

tradional organizational structure, can often drain limited resources and result in a lower standard of quality for a science gateway.

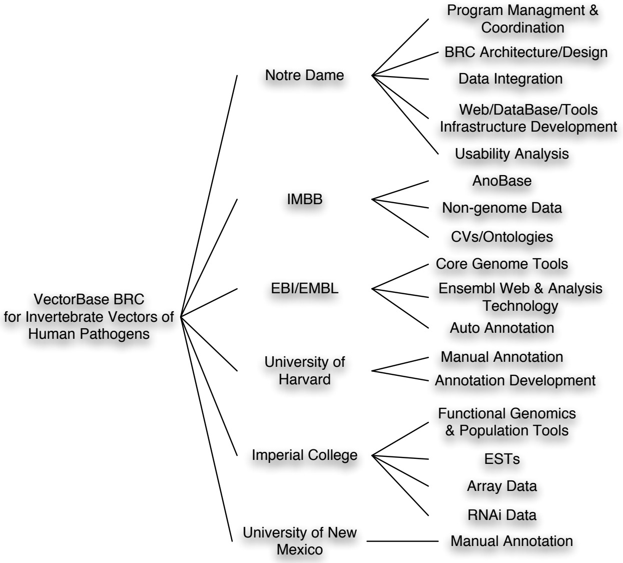

Fig. 1. The organizations and their primary areas of responsibility, for contributing to the VectorBase scientific gateway.

To address these challenges, the VectorBase team has developed and uses a project management cyberinfrastructure to facilitate synchronous and asynchronous distributed collaboration. Those services include, but are not limited to, e-mail and mailing lists, electronic chat, issue trackers, voice and video over conferencing, project management and documentation wikis, a content management system (CMS), and source code versioning and control systems. This project management cyberinfrastructrue employ commercial software packages, like Jira for issue tracking, and, where possible, free and open-source software packages. By integrating these solutions, the resulting cyberinfrastructure allows the BRC to realize the benefits of a VO and overcome the challenges of both VO and non-VO structures and fine-tune the utilization of limited resources.

III. Integrating Data and Tools

The community the BRC services has diverse needs and requirements. For some community members, an analysis workflow integrating the ever-increasing amount of vector-related genomic data with the latest analysis tools is a necessity. Other members seek authoritative ontologies or training materials. Others still, simply need a visual reference for the vectors that are the subject of study.

To address these and future needs, the BRC was designed to be modular and extensible. VectorBase system software is modular, built from a variety of open source software tools, bioinformatic systems, and applications. These tools are grouped into four categories: analysis, search, data browsers, and integration & pipelines as shown in Fig 2. Integrating such a diverse set of tools is a daunting, but typically necessary, challenge for science gateways. This challenge becomes even more evident when considering the underlying technologies these tools depend upon (See Fig. 3). VectorBase is both building and maintaining a biological cyberinfrastructure to overcome this challenge.

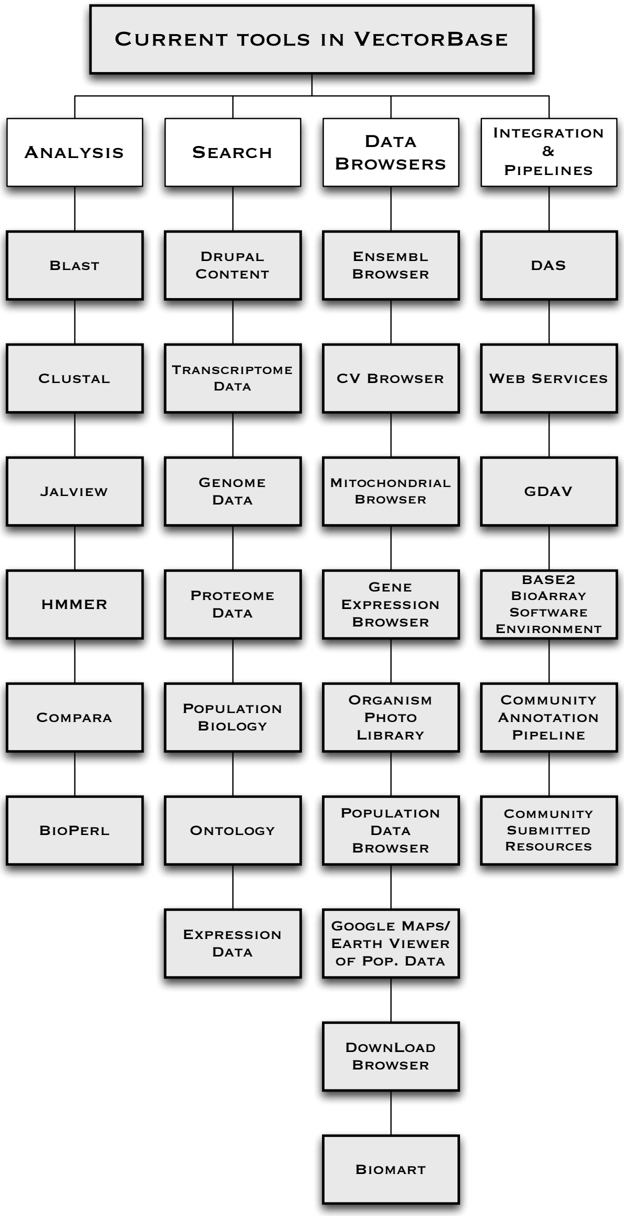

Fig. 2. The major tools deployed within the VectorBase scientific gateway.

A key component of this cyberinfrastructure is the use of the Drupal CMS. Many of the tools and service shown in Fig. 3 are embedded within Drupal by creating custom modules that mediate communication from the CMS (primary means of user interaction) to the tools (performing the desired work). Using a customizable CMS provides a consistent user interface and experience (UI/UX) to the community regardless of the underlying technology required by the suite of tools and services provided. Furthermore, The Drupal CMS can be extended with a variety of independently developed and freely available modules to provide additional functionality without having to invest the resources needed to develop this functionality from the ground up. This approach has afforded the BRC team a high level of agility and flexibility to focus resources on the needs of our community.

An additional feature of the BRC cyberinfrastructure is its use of platform virtualization. Although we employ many physical servers from a variety of vendors, most of the services we provide are contained within independent virtual machines distributed across these physical resources. By virtualizing these services, we can maximize the physical resources by customizing how much each virtual service is allocated. This also provides the flexibility to extend into cloud-based service as needed. As was discussed with respect to establishing a VO, cyberinfrastructure plays a critical role here to maximize resource utilization efficiency.

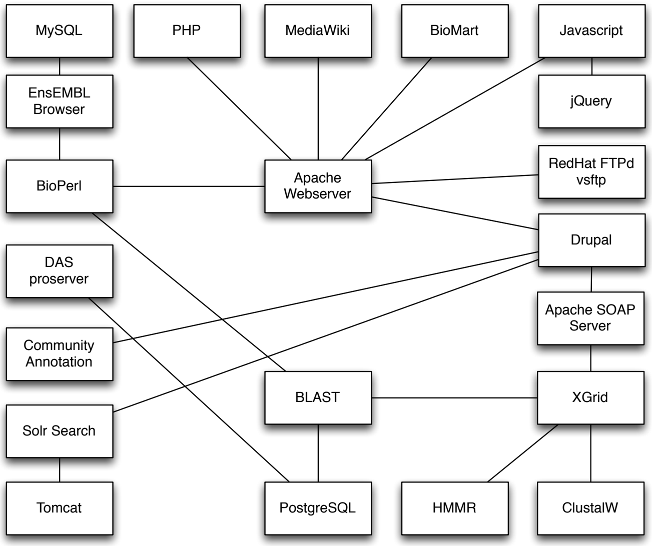

Fig. 3. The major system software components that the VectorBase scientific gateway is assembled from.

IV. Community Involvement and Outreach

Without a sufficiently sized, and sustainable community involvement, a scientific gateway can wither and fade into obscurity. Thus, community involvement and outreach is a critical component necessary to ensure the success of a scientific gateway. The BRC has undertaken a number of initiatives facilitate community involvement.

For instance, the BRC partnered with subsections of our community to help manage the growing amount of data coming from cheaper and widely available genetic sequencing techniques. The biological cyberinfrastructure was extended to include a community annotation portal (CAP) whereby curatorial duties (such as genome annotation) can be performed by individual groups and preserved for consumption by the larger community. This resulted in functionality that is now in use by three other projects [1].

Furthermore, the BRC provides or participates in workshops and conferences relevant to the communities it serves. In a recent arthropod genomics symposium, 50 people were instructed on the use of the BRC. This experience not only provided the community with education on using the site's tools and resources, but also informed the BRC on how they wanted to use the site and what features and changes would improve its usability. Furthermore, these social connections typically result in an increase of community-submitted data.

These community interactions present the opportunity for a synergistic effect. The gateway is continually improving in response to the community's needs, which in turn increases the community's utilization of the gateway. This increase in utilization comes not only from the existing user-base, but also from new community members as the material presented in the workshops becomes used as training material in the classroom. The cyberinfrastructure provides a key role in keeping community involvement high by automating this feedback loop.

V. Discussion

Science gateways offer the promise of developing a critical mass of data, tools, and community members to facilitate scientific research. In developing the VectorBase BRC, three key factors were identified as contributing to the success of this science gateway: virtual organization, data and services integration, and community involvement and outreach. Developing the necessary and appropriate cyberinfrastructure was critical to ensure these three factors contributed to a successful science gateway.

Acknowledgment

The VectorBase Consortium is composed of: European Bioinformatics Institute, UK: Mikkel Christensen, Paul Kersey, Nick Langridge, Daniel Lawson, Gareth Maslen. Harvard University, USA: David Emmert, William Gelbart, Susan Russo, Pinglei Zhou. Imperial College London, UK: George Christophides, Robert MacCallum. Institute of Molecular Biology and Biotechnology, Crete: Emmanuel Dialynas, Vicky Dritsou, Christos Louis, Elvira Mitraka, Pantelis Topalis. University of New Mexico, USA: Phillip Baker, Margaret Werner-Washburne. University of Notre-Dame, USA: Antelmo Aguillar, Steve Bogol, David Campbell, Frank Collins, Katie Cybulski, Scott Emrich, Gloria Giraldo-Calderon, Jarek Nabrzyski, Caleb Reinking.

References

[1] Megy, Karine, et al. "VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics." *Nucleic acids research* 40.D1 (2012): D729-D734..

[2] Lawson, Daniel, et al. "VectorBase: a data resource for invertebrate vector genomics." *Nucleic acids research* 37.suppl 1 (2009): D583-D587.

[3] Lawson, Daniel, et al. "VectorBase: a home for invertebrate vectors of human pathogens." *Nucleic Acids Research* 35.suppl 1 (2007): D503-D505.

[4] VectorBase, 2013, http://www.vectorbase.org