CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2011; **23**:269–278
Published online 2 December 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.1685

Portals for collaborative research communities: two distinguished case studies

Ibrahim Elsayed[1,*,†], Gregory Madey[2] and Peter Brezany[1]

[1]*Department of Scientific Computing, Faculty of Computer Science, University of Vienna, Nordbergstrasse 15/C/3, 1090 Vienna, Austria*
[2]*Department of Computer Science and Engineering, College of Engineering, University of Notre Dame, 384 Fitzpatrick Hall, Notre Dame, IN 46656-0369, U.S.A.*

SUMMARY

Case study research excels at bringing us to an understanding of a complex issue or object and can extend experience or add strength to what is already known through previous research. The research work summarized by this paper discusses two different case studies in the field of portals for collaborative research communities, in particular *VectorBase* and *BGA-Space*. *VectorBase* at its core is a scientific database that focuses on search, data mining and offers multiple integrated bioinformatics tools for analyzing and browsing genomic and related data. *BGA-Space* focuses on capturing semantics from scientists during processing of scientific experiments as well as preserving the full life cycle of scientific data to enable their reuse. The two case studies involve heavy research and the application of theories, concepts, and knowledge commonly discussed in the targeted field. Copyright © 2010 John Wiley & Sons, Ltd.

Received 11 March 2010; Revised 22 September 2010; Accepted 26 September 2010

KEY WORDS: research community portals; e-Science; VectorBase; BGA-Space; scientific dataspace

1. INTRODUCTION

Cyberinfrastructures for e-Science [1] promise to change the way scientists will tackle future research challenges in a number of domains, including earth sciences [2], medicine [3], and life sciences [4]. Great progress has been made in the last decade to utilize service-oriented architectures (SOA) [5] in Grid computing [6] in order to facilitate the virtualization of heterogeneous resources, such as data sources and computational resources. An e-Science portal is a conventional web portal that sits on top of a rich collection of web-based services that allow a community of users access to shared data and application resources without exposing them to the details of Grid computing [7]. Owing to wireless connectivity improvements and hardware getting mobile and constantly smaller and cheaper, portal developers are facing new challenges. Enormous amounts of data will be produced at a rate never seen before in any field of human activity, requiring next generation e-Science portals to cope with and making use of it. Social networking tools are being intensively used by scientists forming virtual scientific communities. This led to an evolution of digital scientific discourse [8] and other dynamics that drive virtual scientific activities, such as research intelligence [9] and workflow-using e-scientists [10], in a way making it important to

*Correspondence to: Ibrahim Elsayed, Department of Scientific Computing, Faculty of Computer Science, University of Vienna, Nordbergstrasse 15/C/3, 1090 Vienna, Austria.
†E-mail: elsayed@par.univie.ac.at

preserve scientific experiments on the whole, including primary data, intermediate data, derived data, the processes, the tools, and their versions used. In this context e-Science portals providing tools to enhance collaboration of scientists are gaining more and more attraction within various research domains.

On the one hand an e-Science portal can provide the user with a single point of access to information, data, and tools that is available and maintained in some kind of organized and distributed scientific space. On the other hand, by having the scientists in front of the portal conducting scientific experiments, portals can also be utilized as instruments to capture information about what the scientist is doing, why, and in what context. Once this crucial semantics about scientific experiments has been organized in an efficient manner and attached to its corresponding primary and derived data, they can provide deeper insights into studies than could be grasped from publications or technical reports.

In this paper we highlight the common problems in the field of portals for collaborative research communities and illuminate those problems through the in-depth study of its application to two different user communities in the life science domain. The paper is structured as follows. The challenges addressed in VectorBase and BGA-Space are discussed in the next section. Before we compare and evaluate VectorBase and BGA-Space in Section 5 we present the two portals, their user communities, and their main services individually in Sections 3 and 4, respectively.

2. CHALLENGES ADDRESSED IN VECTORBASE AND BGA-SPACE

e-Science [1, 11] and cyberinfrastructure [12, 13] programs are initiatives focused on re-energizing and expanding the use of the web and related services to enable more effective research, global collaborations, better utilization of unique resources, and to help address emerging challenges to scientific research. The initial focus was on Grid, distributed, and high performance computing. Problems of workflow, provenance, middleware, and interoperability were addressed early on. A 2007 NSF report added (1) data, data analysis, and visualization; (2) cyber services and virtual organizations (VOs); and (3) learning and workforce development to the goals of the cyberinfrastructure vision [13]. The first item recognizes that science is becoming increasingly data-driven as low-cost sensors, low-cost storage, faster networks are enabling the construction of large data archives that in turn permit discovery through data mining. This is exactly the situation with VectorBase, as second- and third-generation sequencing technologies are producing massive amounts of genomic and derived data. In BGA-Space, similarly, the amount of breath gas source data is increasing enormously due to the high availability of modern analytical instruments such as mass spectrometers. These large data require computational pipelines for feature discovery and data curation. The second item above from the NSF report recognizes that science is increasingly conducted by larger teams (big science), requiring researchers with specialized skills not always locally available, resulting in distributed virtual teams. Concurrently, a new generation of scientists has grown up with the web and social media and are comfortable and proficient with cyber services. This is also a factor influencing the design, development, and usage of both portals; the VectorBase team is distributed across the US and the EU, the scientific users are often younger graduate students and post-docs from over the world, and expectations are high for usability, and contemporary Web 2.0 cyber services. Breath gas analysis researchers form together a large community with members spread over the world. BGA-Space is going to be the key turning point for their collaboration and knowledge exchange. Therefore, expectations for usability and cyber services are very high, similar to VectorBase. Finally, the third item in the NSF report recognizes that an important role for cyberinfrastructure is education and workforce development. Members of the VectorBase team periodically provide training courses at workshops and conferences on the use of the portal, tutorial materials are developed and posted online, and services to enable collaboration and end-user support of other users. Although BGA-Space is in an earlier development phase compared to VectorBase there are strategies for dissemination that include educational activities

DOI: 10.1002/cpe

for instance, tutorials and training courses at meetings of the International Association for Breath Research.

In all of the above aspects VectorBase (already moved into production) provides useful information and tips for the ongoing development of BGA-Space, which is in an early stage of deployment. However, BGA-Space addresses specific challenges that are unique in the field of portals for collaborative science. BGA-Space categorizes their breath gas-related scientific data into three major categories: (1) primary data, (2) derived data, and (3) background data. Data items of each category interact within scientific experiments. Typically, a primary dataset is accessed and analyzed by various analytical methods, which produce a set of derived data products. Analytical methods are often composed into scientific workflows represented by a workflow language and executed by a workflow enactment engine. Background data typically represents such scientific workflows, but is not limited to it. Any data item corresponding to a scientific experiment that does not represent primary or derived data, is classified as background data in BGA-Space. Managing the results of scientific experiments in conjunction with its corresponding input data by enriching the existing relationship among primary and derived data with semantics to be searchable represents an open research faced by BGA-Space. The distributed space of primary, background, and derived data can, if enriched by semantically rich relationships among participating data items support scientists in organizing and preserving their scientific experiments in the long-term to be re-used by owners and others. Thus the challenges we face with the semantic enrichment of data being involved (including scientific discourse) in a scientific experiment are key to the realization of portals for collaborative research communities. Therefore, research challenges addressed by BGA-Space include (a) to interconnect the abovementioned three data categories and to semantically enrich the relationships among them; (b) to invent a suitable relationship paradigm for the creation, representation, and advanced searching of relationships among data of scientific experiments; and (c) to address the full life cycle of data to achieve well-preserved data about scientific experiments conducted.

3. VECTORBASE: BIOLOGICAL CYBERINFRASTRUCTURE FOR RESEARCH ON INVERTEBRATE VECTORS OF HUMAN PATHOGENS

VectorBase (http://www.vectorbase.org) is a bioinformatics portal that focuses on storing genomic and related data on invertebrate vectors that transmit human diseases [14]. Development started in 2004 with sponsorship from the US National Institute of Allergy and Infection Diseases (NIAID). It is one of four such portals, called Bioinformatics Resource Centers (BRC) by NIAID, with the objective of providing free and publicly available web-access to data, bioinformatics tools and services for use by the scientific community. As its name suggests, VectorBase is a database archiving a variety of data types, including: genomic, microarray, gene expression, EST, images, spatial prevalence of genetic varieties, images, and genotype/phenotype association data. VectorBase is managed, developed, and maintained by a distributed team of biologists, bioinformaticians, and computer scientists, with team members located in several locations in the US and the EU.

The distributed team of over 20 persons, both part-time and full-time, is responsible for cyberinfrastructure development, data curation, generation of derived data, end-user outreach, training and support. The scientific user community is world-wide with several thousand unique users per month visiting and using resources provided at the portal. The vectors whose data are archived in VectorBase include multiple mosquito species, a tick species, a louse species, and a multiple species of flies. These vectors transmit diseases by passing pathogens (viruses, bacteria, and parasites) from human-to-human while biting for blood meals. More than a million persons die each year from these diseases, such as malaria, yellow fever, dengue fever, lyme disease, Chagas disease, typhus, leishmania, sleeping sickness, and lymphatic filariasis. Hundreds of millions more are infected each year and, although not fatal, suffer from these debilitating diseases. The data are used to improve the understanding of the vector's biology to enable the development of new and

DOI: 10.1002/cpe

better public health interventions, such as vector control measures, new chemicals to counter the emergence of insecticide resistance in the vectors, new pharmaceuticals, and even new tools for genetic manipulation.

3.1. Biological data in VectorBase

The primary data in VectorBase comes from external sources that sequence vector genomes of interest, i.e. identifying the long sequences of nucleotides that comprise a genome. Sequencing was an expensive, time-consuming process performed only by major genomic centers, such as the Welcome Trust Sanger Institute in the U.K., the J. Craig Venter Institute, the Broad Institute, the Genome Center at Washington University, and others. Next generation high-throughput sequencing technologies are lowering the cost and reducing the time to sequence, thus providing VectorBase with a rapidly growing amount of data from new vector species. After the data are sequenced and assembled into complete genomes, it is processed to discover the location and identity of genes and other biological features. These feature descriptions, meta-data called annotations, are stored in association with regions of the genome. Other experimental methods, such as microarray experiments, contribute additional data describing new features and functions of the genome. Although the VectorBase team contributes new meta-data through computational bioinformatics, the primary data and a growing amount of annotations comes from the research community. Other data in VectorBase includes reports, citations, images, and links to other resources. These data are growing rapidly, and new data types are emerging for an expanding list of vector organisms.

3.2. Cyber services in VectorBase

The cyber services in VectorBase address the needs of researchers at two levels: (1) at the bioinformatics level, helping them utilize data stored in VectorBase, and (2) at the collaborative, user support, and communication levels, enabling researchers to interact with other researchers and the VectorBase staff. Bioinformatics tools that execute on VectorBase servers include: data browsers, search tools, pattern matching tools, browsers of semantically organized controlled-vocabularies, interactive maps displaying genomic variation data, and services for downloading and uploading data. Collaborative, support and communication tools include over 25 specialized mail lists, discussion forums, help and documentation wiki, newsletters, frequently-asked-questions (FAQs), interactive Web 2.0 style tool tips, and online tutorials. These latter services contribute to learning, increased research productivity, better utilization of the bioinformatic resources and workforce development. Cyber services are also used by VectorBase team as described in the following section.

3.3. The VectorBase virtual organization

The VectorBase project is supported by a VO. The VectorBase team is composed of principle investigators, managers, bioinformaticians, developers, computer scientists, and biologists. The team is ‘virtual’ because it is dispersed globally; it is an ‘organization’ because the team has common shared goals. In a publication by the National Science Foundation we see the following:

> A VO is created by a group of individuals whose members and resources may be dispersed globally, yet who function as a coherent unit through the use of cyberinfrastructure (CI). [They] extend beyond small collaborations and individual departments or institutions to encompass wide-ranging, geographically dispersed activities and groups. This approach has the potential to revolutionize the conduct of science and engineering research, education, and innovation [15].

Cyber services are employed by the VectorBase team to enable synchronous and asynchronous distributed collaboration. These services include e-mail, electronic chat, issue trackers, voice and video over IP, a project management wiki, a documentation wiki, and teleconferencing for groups

DOI: 10.1002/cpe

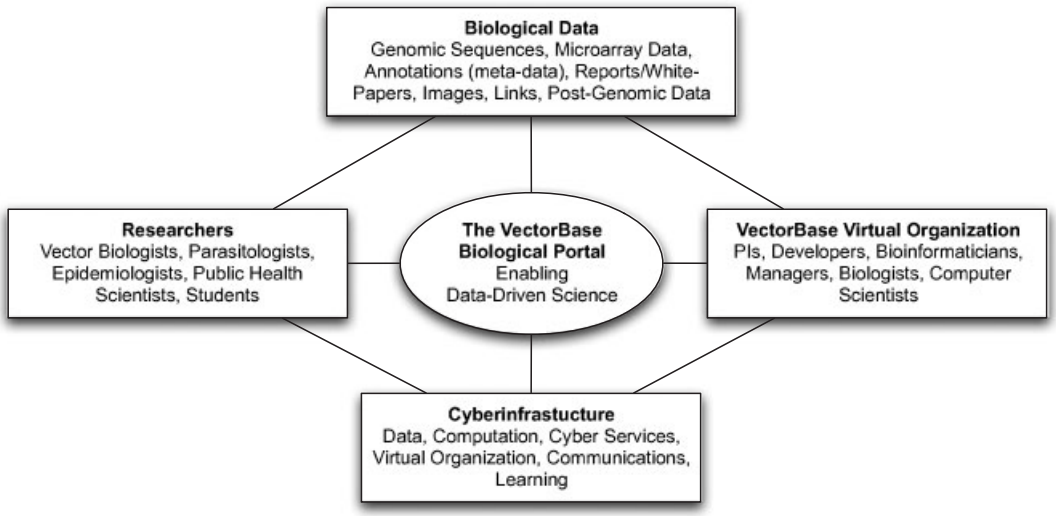

Figure 1. The VectorBase biological portal enables data-driven science.

as large as 25 participants. The VectorBase VO is both building a biological cyberinfrastructure and using cyberinfrastructure to enable its effective performance (see Figure 1).

4. BGA-SPACE: SCIENTIFIC DATASPACE PORTAL FOR THE BREATH GAS ANALYSIS RESEARCH COMMUNITY

BGA-Space is a first prototype of a web-based portal for the breath gas analysis research community. Breath gas analysis is an emerging new scientific field with a growing international scientific community addressing many different breath gas studies in terms of investigating and screening for hundreds of compounds in exhaled breath gas. There is strong evidence that specific cancers can be detected using the concentration pattern of volatile compounds in exhaled air. The purpose of *BGA-Space* is to enable collaborating scientists and institutions several important activities, which include: (a) access to distributed breath gas data and analytical resources collected and developed at different research institutions around the world and (b) to easily contribute to and leverage the resources of an international- and national-scale, multi-institutional environment. This will strongly support global collaborations of scientists, improve decisions, and increase the chance and scope of discoveries in the breath gas research domain.

BGA-Space is built on top of *jSpace* [16]—a *Scientific Dataspace Support Platform* (SDSSP), which we define as a set of software programs that control the organization, storage, and retrieval of data in a distributed space of data [17]. Our data management approach is based on the Grid and *dataspace* concepts. The idea of a future data management paradigm called *dataspace* was introduced by Franklin *et al.* [18] and also addressed by the authors of this paper in [17, 19]. The goal is to manage a dataspace, rather than a database or other dataset types. Dataspaces are modeled as participants (datasets) and relationships. The concepts of a scientific dataspace paradigm are described in [19]. It introduces a specific model of the e-Science life cycle. jSpace realizes an SDSSP on top of Semantic Grid [11] technology. It aims at providing associated mechanisms for managing semantically rich relationships among scientific data sources (primary data) and its corresponding findings (derived data). The latter result from a set of activities defining concrete preprocessing and analysis methods (background data) that were applied to a source dataset. A first prototype of BGA-Space is deployed to a leading breath gas analysis research group acting as a small core of early adopters that will provide us with important feedback and drive our research and plans for further implementations. The conception and prototypical implementation of the scientific dataspace paradigm for breath gas analysis is described in [20].

DOI: 10.1002/cpe

4.1. Preservation of BGA experiments

Breath gas *source data* are fundamental for simulation and modeling by the acting research group. The output of breath gas analyses aims at defining a large number of predictions and might provoke further experimentation, which in turn may take days or weeks, depending on the computational and human resources available. However, the resulting *derived data*, that have arisen from the research task represent valuable information not only to the acting research group, but also to other groups with respect to other main focuses. In this context there is a need for a supporting information infrastructure accessible through an easy-to-use portal providing advanced data management features that enable breath gas analysts (a) to keep track of their e-Science activities and (b) to publish results of breath gas analysis experiments linked together with their source data and well-defined semantics. This is challenged by the BGA-Space portal and its underlying scientific dataspace.

Our approach is to preserve both relationships and data together within the dataspace to be reused by owners and others. To enable their reuse, data must be well preserved. The effects of data loss can be economic, because the experiments have to be re-run, but in some cases data loss represents an opportunity lost forever [21]. Preservation of scientific data is therefore a major requirement, which can best be established if the full life cycle of data is addressed. This is achieved in our approach by the e-Science life cycle model [19], which classifies on a high level of abstraction the steps a scientist is conducting into five major activities (*GoalSpecification, DataPreparation, TaskSelection, TaskExecution, ResultPublishing*), which we call the e-Science life cycle activities. The e-Science life cycle ontology [22] defines the concepts of the e-Science life cycle model as OWL-classes and properties. In cooperation with leading breath gas researchers we have defined in [20] a number of actions that a scientist conducts during the process of performing breath gas studies, which we then mapped to activities of the e-Science life cycle. Based on this common understanding we have developed the first prototype of BGA-Space.

4.2. The BGA dataspace

The breath gas analysis scientific dataspace consolidates four major kinds of interconnected data bases: (1) primary databases for storing input datasets, (2) background databases for storing analytical methods used to analyze an input dataset, (3) derived databases for storing results of analyses tasks, and (4) RDF databases for storing instances of the e-Science life cycle ontology. Figure 2(a) illustrates the main entities of a typical breath gas analysis experiment and (b) shows

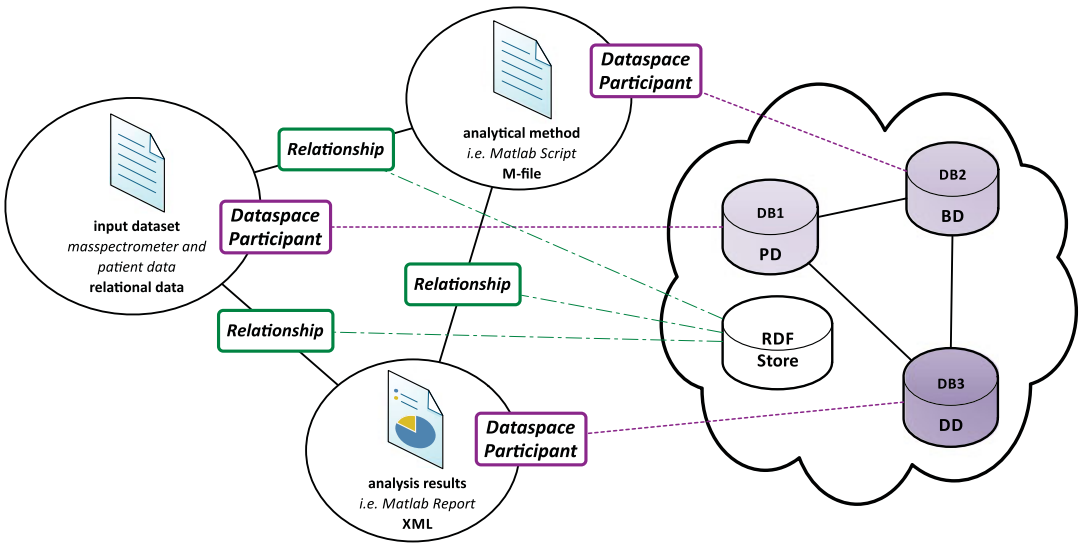

Figure 2. Dataspace participants and relationships in BGA-Space: (a) breath gas experiment and (b) scientific dataspace.

DOI: 10.1002/cpe

their corresponding data bases and how they are organized in the scientific dataspace. An important point here is the RDF-store containing relationships among primary, background, and derived data items that participate in breath gas experiments in terms of individuals and properties of the e-Science life cycle ontology. The dataspace as depicted in Figure 2(b) represents an instance of a breath gas analysis dataspace like it is deployed as an experimental framework for the Breath Research Institute of the Austrian Academy of Sciences. A large-scale scientific dataspace scenario with multiple geographically distributed data bases is elaborated in [16]. Our main contribution in this context is the creation of semantically rich relationships among data items of scientific experiments coordinated by the BGA-Space portal. Based on the relationships answers to specific questions, such as the following:

A '*I have detected a model error and want to know which derived data products need to be recomputed*'.
B '*I want to check if inspiration is different to expiration of breath gas dataset x. If the results already exist, I'll save hours of computation*'.
C '*Is there any experiment done on the volatile organic compound isoprene on exhaled breath gas in the context of cholesterol level in blood*?'

can be answered by submitting SPARQL queries to the RDF-store, which manages the relationships.

4.3. The BGA-Space portal

Breath gas researchers interact with the scientific dataspace via the community web portal. It provides the necessary services to compose, access, and publish breath gas experiments as well as to search, query, and browse the scientific dataspace.

Experiments on exhaled breath gas are being successively refined, by the acting researcher until the study either shows a significant result (i.e. definition of accurate methods for estimation of blood gas levels of certain biomarker values from breath gas samples) or ends up in a modification of the intended defined goal specification for that experiment. We are aware that we rely on active participation of members from the scientific community in order to establish a large-scale scientific dataspace for breath gas analysis. Therefore, we provide a portal with a simple interface that can easily be used by scientists from diverse research domains, especially by non-computer scientists, which was a major requirement from our driving application. However, we suspect that young researchers (Master and PhD students) will be the major users who will use the portal in terms of conducting experiments, while senior researcher will most likely interact with the portal in terms of submitting requests. Currently, we improve the BGA-Space according to the feedback received from the small core of early adopters. Once BGA-Space is deployed to the global breath research community we expect that it enforces building of collaborations among breath gas research institutions as it supports the community in exchanging data and knowledge. This will build the basis for automation-based breath gas analysis, which is one of our future targets.

5. COMPARISON AND EVALUATION OF THE TWO PORTALS

The two collaborative portals, BGA-Space and VectorBase, both have as a focus the support of a distributed research community through online databases and research support tools. The two portals have many similarities, but also differ in many respects resulting in different designs and different measures of effectiveness. In this section, we discuss their similarities and differences, and how those differences influenced their design and development. Additionally, we describe the methods used to evaluate their designs and effectiveness, along with the development challenges and solutions.

5.1. Comparisons

BGA-Space is still under development with a small number of test users; VectorBase development started in 2004 and was deployed as a prototype about two years later. Since then VectorBase

DOI: 10.1002/cpe

has moved into production but development is ongoing, continuously adding new features and new data. The two research communities served by the portals are of different sizes. BGA-Space supports a research community of several hundred users; VectorBase is used by a community of several thousand users. The portal development teams are also different in size with BGA-Space supported by several centrally located developers versus a distributed development and data curation team of approximately 25 contributors on VectorBase (although about half are full-time, the rest part-time). These differences in sizes of the user communities, deployment stages, and the sizes and locality of the development and curation teams result in different emphases in the designs of the portals. VectorBase, because it is already deployed to a larger user community must more carefully ensure the uptime of the servers used by the researchers; this requires the added complexity of multiple instances of the portal: (1) a stable production version that does not change often and must support many concurrent users, (2) a pre-production version with new features and data which is undergoing testing and final refinements prior to moving into production, and (3) development versions of the portal with incomplete and buggy features that are not yet ready for migration to the pre-production version of the portal. The smaller and more centrally located BGA-Space development team does not require elaborate coordination and collaboration tools; while the somewhat larger and distributed VectorBase team requires multiple collaboration systems, such as developer wikis, multiple mailing lists, instant messaging, video and screen sharing chat services, and large telephone conference calls.

Both portals use ontologies to help improve data management and use, but toward different goals. BGA-Space utilizes the e-Science life cycle ontology [22] to add rich semantics to data coming from scientific experiments. Thus, BGA-Space provides the recently introduced neighborhood keyword queries [23] and advanced semantically rich queries against the databases. For example, searching for 'endogenously-derived gases' returns not only the *goalSpecification* individuals available in the dataspace that mention 'endogenously-derived gases', but also those of its neighbor activities informing the user what proband data was used and where it resides, which analytical methods were applied, and what results of the corresponding experiment were achieved. Information about the scientist who conducted the experiment is also returned in this context. VectorBase, an older system, is less advanced in its use of ontologies, with their primary purpose to provide standard terms for search queries and meta-data annotations.

Another difference between the two portals results from the sources of the data in the portal's databases. BGA-Space's data comes from experiments coordinated by the portal and related systems; thus, the BGA-Space user community are the originators of the data. BGA-Space organizes the data into three different types of databases (primary, background, and derived databases) and interconnects them by creating semantically rich relationships among them. The data in VectorBase come from external sources, such as the large sequencing centers at the Welcome Trust Sanger Institute or the Broad Institute, from researchers performing gene-expression experiments using GeneChips, and many other experimental methods. VectorBase serves as a centralized archive for such data, but generally does not support the workflows that generate these data.

5.2. Evaluation

Although there are differences between BGA-Space and VectorBase, similar methods can be used to evaluate the portals' effectiveness and performance. The ultimate measure of effectiveness is whether the target research community uses the portal, and reports satisfaction with its services. Several indirect measures of this usage and satisfaction are used. BGA-Space, a portal still in the prototype stage, is evaluated by a small core of early adopters. Feedback tends to focus more on usability, bug reports, feature requests, etc. VectorBase, already deployed with several thousand unique users each month, collects usage statistics using Google Analytics and Apache web server log analysis. User forums (i.e. electronic bulletin boards) collect complaints, bug reports, feature requests, and through the FAQs, areas for improvement. As part of an outreach program to publicize the availability of the VectorBase portal, tutorials and posters at conferences and workshops are offered, providing an opportunity to engage the users. Some surveys of users have been conducted at these conferences and workshops attended by the VectorBase research community. Typical

DOI: 10.1002/cpe

feedback includes feature requests (new tools and new data) and improvements to usability. We expect that as BGA-Space moves into production, similar measures of usage and user satisfaction can be applied. For example, BGA-Space has been presented to the community at a BGA scientific conference.

5.3. Technical challenges

The VectorBase project had multiple challenges in several areas. First, since it is primarily a large complex bioinformatic database, the design of its database schemas was a challenge given that several standards exist with their own respective suites of existing compatible software tools. The solution was to employ two standards (Chado and Ensembl), providing access to their suites of tools, but resulting in extra work maintaining partially redundant copies of the data. A second related challenge was the integration of a large number of existing open-source tools. Techniques such as XML-based data interchange and software wrappers provided solutions. The third challenge was the coordination of the distributed team of developers and scientists; this was addressed by the tools described earlier including wikis, mail lists, instant messaging, periodic conference calls, and IP-telephony.

One of the main technical challenges we faced in BGA-Space was to keep the e-Science life cycle ontology and also the web-portal domain independent. Currently, the tools being developed can easily be ported to another application domain.

6. CONCLUSIONS

This paper presented VectorBase and BGA-Space, two distinguished case studies in portals for collaborative research communities. VectorBase at its core is a scientific database offering data mining and multiple integrated bioinformatics tools for researchers studying invertebrate vectors that transmit human diseases. BGA-Space addresses the preservation of breath gas analysis studies to enable their reuse. Besides that both portals are implemented for different research communities, their underlying infrastructures are different in the way they organize scientific data. The motivation for this joint paper was to provide an elaboration of two independent approaches that are in the scope of portals for collaborative research communities. Both issues were raised at the International Workshop on Portals for Life Sciences (IWPLS'09) hosted by the e-Science Institute in September 2009. As a result we see in our future work both, an integration of dataspace services into the highly visited bioinformatics resource portal VectorBase as well as the utilization of VectorBase data mining tools for the breath gas analysis research community.

ACKNOWLEDGEMENTS

We thank Sandra Gesing and Jano van Hemert for inviting us to contribute to this Special Issue on IWPLS'09. The VectorBase Bioinformatics Resource Center is supported by the National Institutes of Health—National Institute of Allergy and Infectious Diseases under Contract Number HHSN272200900039C. The Austrian BMWF (Federal Ministry for Science and Research) funding of the Austrian Grid 2 project (Contract: GZ BMWF-10.220/0002-II/10/2007) is key to bringing BGA-Space partners together and to undertaking the research.

REFERENCES

1. Hey T, Trefethen AE. Cyberinfrastructure for e-science. *Science* 2005; **308**(5723):817–821.
2. Ramakrishnan L, Simmhan Y, Plale B. Realization of dynamically adaptive weather analysis and forecasting in LEAD: Four years down the road. *Dynamic Data-Driven Application Systems Workshop at ICCS*. Springer: Berlin, Heidelberg, 2007.
3. The @neurIST Project. Integrated Biomedical Informatics for the Management of Cerebral Aneurysms. Available at: http://www.aneurist.org [1 June 2010].
4. Krishnan A. A survey of life sciences applications on the grid. *New Generation Computing* 2004; **22**(2):111–126.

DOI: 10.1002/cpe

5. Srinivasan L, Treadwell J. An overview of service-oriented architecture, web services and grid computing. *Hewlett-Packard White Paper*, 2005. HP Software Global Business Unit, 2005 Hewlett-Packard Development Company, V02, 11/2005.
6. Roure DD, Baker MA, Jennings NR, Shadbolt NR. The evolution of the grid. In *Grid Computing*: *Making the Global Infrastructure a Reality*, Berman F, Hey AJG, Fox G (eds.). Wiley: New York, 2003; 65–100.
7. Gannon D, Plale B, Christie M, Huang Y, Jensen S, Liu N, Marru S, Pallickara SL, Perera S, Shirasuna S, Simmhan Y, Slominski A, Sun Y, Vijayakumar N. Building grid portals for e-science: A service-oriented architecture. *Advances in Parallel Computing* 2008; **16**.
8. Ciccarese P, Wu E, Clark T. An overview of the swan 1.0 discourse ontology. *WWW '07*: *Proceedings of the 16th International Conference on World Wide Web*. ACM: New York, NY, U.S.A., 2007.
9. Sheehan J. Research intelligence: A social networking toolset supporting multidiscipliary e-science. *ESCIENCE '08*. IEEE Computer Society: Washington, DC, U.S.A., 2008; 331.
10. Goble CA, De Roure DC. myexperiment: Social networking for workflow-using e-scientists. *WORKS '07*. ACM: New York, NY, U.S.A., 2007; 1–2.
11. De Roure D, Jennings NR, Shadbolt NR. The semantic grid: A future e-science infrastructure. In *Grid Computing—Making the Global Infrastructure a Reality*, Berman F, Fox G, Hey A (eds.). Wiley: New York, 2003; 437–470.
12. Stein LD. Wiki features and commenting—Towards a cyberinfrastructure for the biological sciences: Progress, visions and challenges. *Nature Reviews Genetics* 2008; **9**(9):678–688.
13. *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation, Cyberinfrastructure Council: Arlington, VA, 2007. Available at: http://purl.access.gpo.gov/GPO/LPS80410.
14. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell KS, Christophides GK, Christley S, Dialynas E, Hammond M, Hill CA, Konopinski N, Lobo NF, MacCallum RM, Madey G, Megy K, Meyer J, Redmond S, Severson DW, Stinson EO, Topalis P, Birney E, Gelbart WM, Kafatos FC, Louis C, Collins FH. Vectorbase: A data resource for invertebrate vector genomics. *Nucleic Acids Research* 2009; **37**(Database issue):D583–587.
15. NSF. Engineering virtual organization grants. Available at: www.nsf.gov/pubs/2007/nsf07558/nsf07558.pdf [1 June 2007].
16. Elsayed I, Brezany P. Towards large-scale scientific dataspaces for e-science applications. In *DASFAA Workshops* (*Lecture Notes in Computer Science*, vol. 6193), Yoshikawa M, Meng X, Yumoto T, Ma Q, Sun L, Watanabe C (eds.). Springer: Berlin, 2010; 69–80.
17. Elsayed I, Brezany P, Tjoa AM. Towards realization of dataspaces. *DEXA '06*. IEEE Computer Society: Washington, DC, U.S.A., 2006; 266–272.
18. Franklin M, Halevy A, Maier D. From databases to dataspaces: a new abstraction for information management. *SIGMOD Record* 2005; **34**(4):27–33.
19. Elsayed I, Muslimovic A, Brezany P. Intelligent dataspaces for e-science. *CIMMACS'08*. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, WI, U.S.A., 2008; 94–100.
20. Elsayed I, Ludescher T, Schwarz K, Feihauer T, Amann A, Brezany P. Towards realization of scientific dataspaces for the breath gas analysis research community. *IWPLS'09*: *Proceedings of International Workshop on Portals for Life Sciences*, CEUR, U.K., 2009.
21. Lynch C. Big data: How do your data grow? *Nature* 2008; **455**(7209):28–29.
22. Elsayed I, Muslimovic A, Brezany P. The e-science life cycle ontology. Available at: http://www.gridminer.org/e-sciencelifecycle/lifecycleontology_v2.0.owl [1 June 2010].
23. Dong X, Halevy A. Indexing dataspaces. *SIGMOD '07*: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. ACM: New York, NY, U.S.A., 2007; 43–54.

DOI: 10.1002/cpe