

## A Cross-disciplinary Collaborative Research Platform - Study on Qinghai Lake Joint Research Environment

Juan Zhao

Computer Network Information Center,  
Chinese Academy of Science,  
Graduate University of Chinese Academy of Sciences  
Beijing, China 100190  
zhaojuan@cnic.cn

Shuren Li, Jianjun Yu, Kejun Dong, Kai Nan,  
Baoping Yan

Computer Network Information Center, Chinese  
Academy of Science  
Beijing, China 100190  
lizr@cnic.cn, {yujj, kevin, nankai}@cnic.ac.cn,  
ybp@cnic.cn

**Abstract**—By constructing e-Science projects, we have explored that the creation of a collaborative research platform, is in great demand. This paper explores the requirements and approaches to build a collaborative research platform named Duckling for e-Science projects. Our case study is the e-Science project in joint research center of the Chinese Academy of Sciences and the Qinghai Lake National Nature Reserve, where many scientists from different areas and various backgrounds form virtual teams to study migratory bird protection and avian flu problems. This paper studies collaborative characteristics and requirements of virtual scientific communities. It introduces the framework of Duckling and applies it to the Qinghai Lake case, which helps to construct a real-time collaborative environment for scientists and promotes e-Science activities.

**Keywords**- e-Science, collaborative research platform, information sharing, Web2.0, Grid

### I. INTRODUCTION

With the rapid development of the Internet, and the exceptional increase in computing power, storage capacity and network bandwidth, information sharing and collaboration have widely taken place, especially for scientific researchers. The changing scale and scope of experimental science also require a new research paradigm, i.e. e-Science, where researchers collaborate with each other globally via virtual communities, which is now under construction in many countries for various disciplines [1].

Most of traditional work focused on building cyber-infrastructure to facilitate accessing to the geographically distributed and heterogeneous resources, such as computational resources, scientific instruments, databases, and applications. However they put little concern on communication and collaboration between scientists, which are of great importance in scientific research. For example, when doing research online, scientists may need messaging (email, calendar and scheduling), team collaboration (file sharing, task management), real-time collaboration and communication (application / desktop sharing, voice, audio and video conferencing), and community authorization. Scientists may also need collaboration on data analyzing and sharing. Currently there has some work conducted to build collaborative platforms for e-Science, but functions simply

focus on communication, and lack of support to cross disciplinary research. Communication and collaboration between researchers from various backgrounds become more difficult as they need some common information context for information sharing and require multi-layered and different point of view of data display. Thus, they need a more intelligent collaborative platform to support their online scientific activities within and between virtual communities.

This paper was motivated by the scientific work carried out by joint research center of Chinese Academy of Sciences and Qinghai Lake National Nature Reserve (short for Qinghai Lake Joint Research Base). The lake becomes a focus in global concerns of avian influenza (H5N1), as a major outbreak could spread the virus across Europe and Asia, further increasing the chances of a pandemic. Many scientists from different organizations and various disciplines come to study the bird protection and avian flu problems. In order to support cross-disciplinary, cross-cutting, cross-unit and international collaborative research, we have established a collaborative research platform named Duckling, by means of which, scientists, wherever they are, could carry out research and communicate on line. In this paper, we focus more on how to develop a collaborative environment for real scientific application. We would firstly study on the collaborative characteristics and requirements of the virtual scientific communities in Qinghai Lake Joint Research Base. We would emphasize Duckling architecture which takes advantages of Grid and Web 2.0 technologies making collaborative work online. Furthermore, we would introduce customized realization and effectiveness achieved in the Qinghai Lake case, which helps to construct a real-time collaborative environment for scientists and promotes e-Science activities.

This paper is organized as follows: The second section describes the related work. The third section introduces e-Science work in Qinghai Lake Joint Research Base and summaries the characteristics of collaborations and requirements in cross-disciplinary scientific research. The fourth section introduces the architecture of Duckling and its core components. In section V, we introduce the real applications and collaboration enhancement in the Qinghai Lake case. Section VI concludes this paper and outlines our topics for further research.

## II. RELATED WORKS

We reviewed two lines of work which are closely related to the concept we will propose: e-Science related projects and collaborative application in e-Science.

The TeraGrid [2] integrates high-performance computers, data resources and tools, and high-end experimental facilities. Currently, TeraGrid resources include more than 750 teraflops of computing capability and more than 30 petabytes of online and archive data storage, with rapid access and retrieval over high-performance networks. Researchers can also access more than 100 discipline-specific databases, which make TeraGrid one of the world's largest, most comprehensively distributed cyberinfrastructures.

OSG (Open Science Grid) [3] is a distributed shared cyberinfrastructure for domain scientists, computer scientists, technology specialists, and providers in more than 70 U.S. universities, national laboratories, and organizations providing resources, tools and expertise. The OSG consortium also partners with campus and regional grids, large projects such as TeraGrid, Earth System Grid, Enabling Grids for e-Science (EGEE) in Europe and related efforts in South America and Asia to facilitate interoperability across national and international boundaries.

The Biomedical Informatics Research Network (BIRN)[4] is a geographically distributed virtual community of shared resources offering tremendous potential to advance the diagnosis and treatment of disease. BIRN is changing how biomedical scientists and clinical researchers make discoveries by enhancing communication and collaboration across research disciplines. BIRN hosts a collaborative environment rich with tools that permit uniform access to hundreds of researchers, enabling cooperation on multi-institutional investigations. BIRN synchronizes developments in wide area networking, multiple data sources, and distributed computing.

The VL-e [5] is a Dutch e-Science project. The goal of the VL-e project is to bridge the gap between the technology push of high performance networking of the Grid and the application pull of a wide range of scientific experimental applications. It will provide generic functionalities that support a wide class of specific e-Science applications and set up an experimental infrastructure for the evaluation of ideas. The mission of VL-e is to boost e-Science by creating an e-Science environment and carrying out research on methodologies. Its strategy is to carry out concerted research along the complete e-Science technology chain, ranging from applications to networking, focusing on new methodologies and re-usable components. The essential components of the total e-Science technology chain are: 1) e-Science development areas, 2) a Virtual Laboratory development area, 3) a large-scale distributed computing development area, consisting of high performance networking and grid parts.

The "CyberBridges" project [6] is a model collaboration infrastructure for e-Science, designed to complement the technical infrastructure with a socio-organizational service-oriented infrastructure. It offers the potential of creating a community of scientists and researchers capable of

collaborating with their counterparts across the network by fully integrating cyber-infrastructure into their educational, professional, and creative processes.

## III. E-SCIENCE IN QINGHAI LAKE JOINT RESEARCH BASE

Our studies were conducted at the Qinghai Lake National Nature Reserve, Qinghai province, China. Qinghai Lake, the largest salt lake in China with an area of 525 km<sup>2</sup>, is located in the middle of Qinghai Province. From 2001, the Chinese Academy of Sciences (CAS) has been aware that e-Science has played a very important role in online scientists' collaboration. CAS initiated the e-Science program in Qinghai Lake Nature Reserve, which is a virtual organization, that many other groups (e.g. Scientific Research Institutes, Qinghai Lake Reserve Government, USGS) utilize. These groups make full use of resources to build small or large research-driven virtual teams to work on migratory bird protection and avian flu problems, such as identification of important migratory bird species, spatial distribution patterns, and disease monitoring and risk assessment.

### A. Characteristics

In this section, we describe the characteristics of e-Science work in Qinghai Lake Joint Research Base that arise from four areas of concern –organizations, research collaboration, resources, and information-sharing.

#### 1) Organizations are research-driven

Like many e-Science projects, Qinghai Lake Joint Research Base was formed for specific scientific research problems with clear overall common goals and objectives in mind. All members have agreed on these goals and objectives and contribute to the realization of the goals.

#### 2) Collaboration runs through whole research activities.

As institutes participating in this program are distributed geographically and the research problems are interdisciplinary, collaboration always runs through all research activities from experimentation, data analysis, and results-publishing and most of the other collaborative activities to IT, technology, and support, such as real-time communication tools, collaborative document editing, and virtual community management.

#### 3) Experiment and the research process are resource intensive.

As scientific activities become larger and more complicated, researchers have to handle larger amounts of data generated in e-Science such as experimental data generated by the migratory experimental instruments, paper data from digital library, documents, reports, patents and so on. As many of the resources are dispersed, researchers needed to locate each resource and query through their own interface, as well as finding the unknown relationship.

#### 4) Information sharing need

Information providers are also in demand thus encouraging information sharing. However, different stakeholders' require different types of information. And people from various disciplines have different perspectives of the same information.

## B. Requirements

Collaborative environments in e-Science are a dynamic mix of organizations (people), research activities, and resources. Current research communities focus on one or, at best, two of the three aspects as described in section II. During our investigation of Qinghai Lake Joint Research Base, there was an urgent need to develop informative and collaborative systems capable of supporting and advancing new modes of research activities. The detailed requirements of the collaborative environment in Qinghai Lake Joint Research Base are:

- Research projects can be used as "virtual organizations" (VO) in the collaborative platform. Researchers could participate in the VO corresponding activities and project, such as meetings, paper reviews, inspections and acceptances, and inspection activities, use the corresponding service and communicate with peers or find a partner.
- Data and information could be shared and reused by different virtual organizations, with a good authentication and authorization management.
- Collaborative platforms should facilitate access to many kinds of heterogeneous resources like experiential instruments, computing, storage, databases, literature as well as domain-specific applications. These resources should not be isolated and their relationship should be uncovered.
- Collaborative platforms should support the ongoing process of scientific discovery and development from experimenting, data analyzing to results publishing.
- The communication such as email, video conferences should also be enabled.
- Researchers share all kinds of electronic documents with projects / virtual partner or other person, in an efficient safe and easy way and writing documents collaboratively.

## IV. ARCHITECTURE OF DUCKLING

We focus more on how to develop a collaborative environment for real scientific application. The implementation of Duckling is based on an architecture that is extensible and supports, but does not restrict, the resources such as databases, instruments and mobile devices required by virtual communities. Fig.1 presents the Duckling architecture. The resources are created by e-Science projects which are widely distributed in different institutes in CAS. Also the types of resources are heterogeneous. The large-scale, distributed resources are universally integrated by a Resource Service in the Duckling architecture. The Resource Service includes different plug-ins aimed to provide pipelines for different types of resources, like Computing Resources Plug-in, AV Plug-in, Device Plug-in, Database Plug-in, Digital Library Plug-in and others. We also provide a scientific workflow over these resources, which can easily draft complex business logic and consume resources to generate scientific knowledge. To help integrate, share and

manage the large scale and distributed resources, we also built three core toolkits for online collaboration, including VO management, document collaboration, and activity collaboration. VO management toolkit is designed for scientific virtual organization management, helping scientists to organize the communities with the same interests. Document collaboration toolkit is designed for document sharing and searching with the support of a text search engine. Activity collaboration toolkit helps to organize Internet-based activities, such as project review and conferences. Also we design a Virtual Work Console to help access, organize and manage resource easily and transparently and integrate resources from CAS, such as computing, storage, database, and digital library. To assist scientists in accessing the resources needed, we provide a CA engine that helps them to use the correct resource.

We will describe core work components in the following sections.

### A. Research and Protocols

As an ultimate goal, Duckling is to help scientists in different domains to manage their complex, long-term and wide-spread research activities online. To realize this goal, we would carry out our research from three aspects:

- Solve the collaboration problem for scientific activities i.e. support the process of the collaborative scientific activities. This includes integrating scientific resources, publishing, creating and maintaining cross-disciplinary and cross-organizational collaboration, and organizing and managing academic conferences.
- Solve the management problem for scientific resources. The scientific resources are distributed and heterogeneous, which also makes the Duckling a distributed environment. We need to provide a loose-coupled architecture that will manage these resources and services, thus supporting the interoperation for upper scientific collaboration.
- Solve the availability and accessibility problem for scientific activities. Scientists can access the collaborative research platform portal to carry out scientific activities anywhere, anytime and on any device.

The first step of research is to investigate the advanced technologies for Duckling architecture, including Grid, Web 2.0, Web services, JSR-268 portal and so on. The architecture must be scalable and extensible enough to make different kinds of plug-ins be loaded easily. The architecture should also enable scientists to create and maintain scientific workflow. We would firstly investigate the requirements of workflow for scientific activities, analyze the workflow like KEPLER, Condor DAGMan, Globus, Pegasus, Chimera, Unicore, Triana and so on, and finally customize a suitable workflow engine and applications for different types of activities.

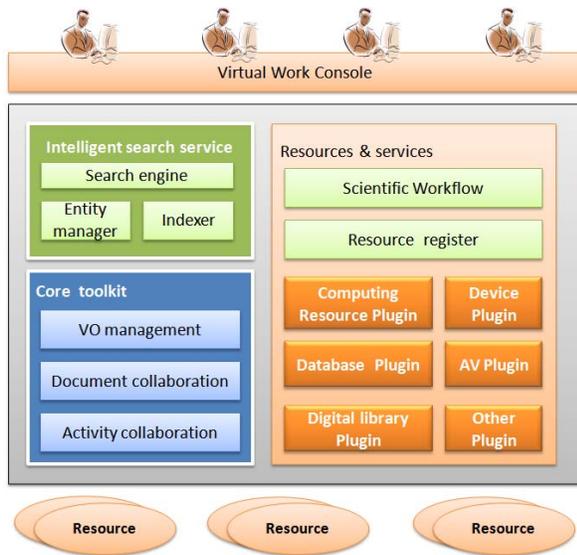


Figure 1. The architecture of Duckling

## B. Virtual Work Console

Virtual Work Console is a portal based Web UI, which provides an EASY-TO-USE interface and a universal access point for different resources and services. It is an open, scalable, flexible integrated platform for different resources based on Portal, Grid and SOA technologies. The Web UI is based on open source portal architecture, which enhances other portals like uPortal, GridSphere, and Liferay. Grid and SOA are used to integrate heterogeneous and distributed resources and provide universal services for upper Web interfaces.

## C. Core Toolkit

### 1) Document Collaboration Tool

Document Collaboration Tool (DCT) is a collaborative writing, document-sharing and management tool for virtual communities, which supports concurrent editing for proposal- planning, schedule tracing, project summary. DCT provides two basic functions: 1) Online collaborative document-writing among group users. Traditional collaborative document-writing is always based on e-mail, which is not real-time and cannot provide document segment modification. DCT is a wiki-based collaborative tool that can support document editing with different granularity. 2) Online document sharing. Scientists can share their scientific documents to members of related communities. We provide a search engine to index the documents for easy searching by other persons. We also provide a semantic mechanism to tag documents from users' viewpoints.

### 2) Activity Tool

Scientific activities not only include the academic conference, but also the out-side investigation, real time observation. Activity Tools (AAT) provides a communication channel from national discussions to international conferences. Activity Tools can easily complete tasks like arranging conferences including calling for papers,

making activities schedules, activities resources sharing, video conference and so on.

### 3) User Management Tool (VO Management Tool)

User Management Tool (UMT) is also a VO management tool which helps to provide a universal VO management and authentication solution. Different applications share the same user system, but provide different privileges for their specific requirements. Through a single sign-on solution, each user can log-in once, and access all applications provided in Duckling. Users are organized in different communities according to their interests, and thus possess their own profiles and access to control. Applications based on VO Management Tool utilize the user/group information to address their specific needs.

## D. Resources Manager

We have implemented JSR286 Portlet framework in the platform - a portlet container to offer an easy deployment environment for plugging-in the portlet application implemented for domain applications. Resources and application plug-ins enrich the platform as a collaborative environment for different requirements while keeping the application autonomous.

## E. Intelligent Search Service

This component supports searching rich content in context, by indexing content, such as data and information, generated by core toolkits or some plug-ins in Duckling, and providing a uniform search interface for searching these contents. Above all, we have implemented an entity management, including an entity recognizer and entity extractor to discover the linkage between the resources and information [7].

## F. Duckling Advantages

Duckling aims to provide a system which presents easier usage than current systems. It provides the following advantages over previous systems:

- Resources integration. Duckling is a resource integrator that provides users with a single operating environment where many kinds of resources, such as supercomputers, mass storage facilities, scientific databases, digital libraries, high bandwidth link, scientific equipment, and etc. could be accessed in a seamless way.
- Customized service. Duckling provides a user with what he or she wants completely and exactly. Each user may have a specific workbench individually. Users may choose different services at different times.
- Ubiquitous research. Duckling benefits from state-of-the-art technologies on mobile computing and related technologies so that users could use the collaborative research platform at anytime and anywhere.
- Collaborative work. Duckling enables a lot of scientists, who are from multiple independent institutions, from multiple sites across the world, and from different professional backgrounds, to work



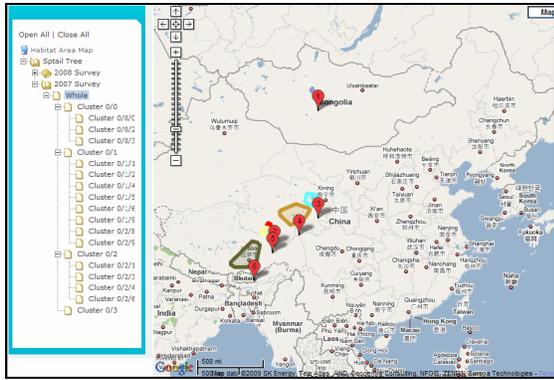


Figure 4. Discovering migration habitats and routes application



Figure 5. Accessing to paper and the relevant data

## VI. CONCLUSIONS AND FUTURE WORK

While more e-Science projects are being implemented, scientific research mode has undergone new paradigms and researchers are more dependent on information and collaborative systems. In this paper, we demonstrate the requirements of collaboration and information-sharing by exploring the frontline e-Science project Qinghai Lake Joint Research Base), which was a typical case of cross-disciplinary, cross-cutting, cross-unit and international collaborative research and poses many difficult problems for enabling researchers to collaborate efficiently. We designed the architecture and realization of the collaborative research platform-Duckling.

As we have implemented and deployed the collaborative platform, currently, we focus our work on developing the following tools to advance the platform.

### A. Semantic Enhancement

We intend to build subject ontology and collaboration ontology to enhance the semantic navigation and search. Currently, we have had user tags, metadata, entities which are all private semantic information, but are not sufficient in

scientific research which needs more accurate and standard semantic information.

### B. Social Network for Virtual Community

Virtual Community is composed of people with the same interests. A social network is a social structure made of nodes (which are generally users or groups in Virtual Lab) that are tied by one or more specific types of interdependency. Research in a number of academic fields has shown that social networks operate on many levels, from families up to the level of nations, and play a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals. So we use social network analysis to show the relationship of users, and discover people with the same interests and so on.

In conclusion, e-Science needs more international collaborations on cyber-infrastructure. E-Science applications are able to merge scientific domains and IT not only in IT technology and scientific knowledge, but also in human, e.g. e-Scientist.

## ACKNOWLEDGEMENTS

This work was supported by Chinese Academy of Sciences (CAS) of Special Items of Information (No.INFO-115 D01); and by Director's Foundation of CNIC, CAS(CNIC\_ZR\_08003);and by Foundation for the Youth of Computer Network Information Center, No.CNIC\_QN\_09003.

## REFERENCES

- [1] Michael Jubb, Keith Adlam, e-infrastructure strategy, Report of the Working Group on Search and Navigation, pp.1, March 2006
- [2] M. Krallinger, A. Valencia, L. Hirschman, Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome biology*, Vol. 9 Suppl 2, No. Suppl 2. (2008)
- [3] W. Cohen and A. McCallum. Information Extraction and Integration: an Overview. In *SIGKDD Conference*, 2004.
- [4] R. Grishman. Information Extraction: Techniques and Challenges. In *SCIE*, 1997.
- [5] <http://www.csdb.cn/>
- [6] Heidi L. Alvarez, David C. Chatfield, Donald A. Cox. *CyberBridges: A Model Collaboration Infrastructure for e-Science*. CCGRID 2007
- [7] Juan Zhao, Kejun Dong, Le Yang, Kai Nan, Baoping Yan, "E-SCHOLAR: Improving Academic Search through Combining Metasearch with Entity Extraction", The 1st IEEE Youth Conference on Information, Computing and Telecommunications (YC-ICT 2009).
- [8] Jiwei Song, Jinyi Wang, Kai Nan, Baoping Yan, Design of Web-Based Remote Monitoring System of Migratory Birds at Qinghai Lake, 2010 Fifth International Conference on Internet and Web Applications and Services.
- [9] Minglie Tang, Yuanchun Zhou, Peng Cui, Baoping Yan: Discovery of Migration Habitats and Routes of Wild Bird Species by Clustering and Association Analysis. ADMA 2009: 288-301.