A Data-Centric Science Gateway for Computational Neuroscience

Shayan Shahand*‡, Ammar Benabdelkader*, Jordi Huguet†, Mahdi Jaghouri*
Mark Santcroos*, Mostapha al Mourabit*, Paul F.C. Groot†
Matthan W.A. Caan†, Antoine H.C. van Kampen* and Sílvia D. Olabarriaga*‡
* Bioinformatics Laboratory, Dept. of Clinical Epidemiology Biostatistics and Bioinformatics
† Brain Imaging Center, Dept. of Radiology
Academic Medical Center, University of Amsterdam, The Netherlands

***Abstract*—Science gateways provide user interfaces and high-level services to access and manage applications and data collections on distributed resources. They facilitate users to perform data analysis on distributed computing infrastructures (DCIs) without getting involved into the technical details. The e-BioInfra Gateway is a science gateway for biomedical data analysis on a national grid infrastructure, which has been successfully adopted for neuroscience research. Necessary improvements in this gateway motivated the design of a new next generation of e-BioInfra Gateway. In this paper we describe the motivation, requirements and design of this new gateway, which is based on the WS-PGRADE/gUSE SG framework, allowing for support for other types of DCIs. The new gateway has additional generic data and meta-data management facilities to access and manage (biomedical) data servers, and to provide an integrated and data-centric user interaction. Its first prototype is implemented and deployed for the computational neuroscience research community of the Academic Medical Center of University of Amsterdam.**

***Keywords*—science gateway (SG), e-science, virtual laboratory (VL), problem solving environment (PSE), computational neuroscience, medical image analysis**

CEUR-WS.org/Vol-993/paper12.pdf

I. Introduction

Science Gateways (SGs), also called Problem Solving Environments (PSEs) or Virtual Laboratories (VLs), support scientists in e-Science endeavours. De Roure et al. [1] described the requirements of e-Science environments as a spectrum with two ends. One end is characterized by automation, virtual organizations of services, and the digital world, and the other end is characterized by interaction, virtual organizations of people, and the physical world. Orthogonal to these requirements at both ends is the issue of scale, for example, of virtual organizations, computation, storage, and the complexity of relationships between them. Increasing scale demands automation and, as highlighted by Hey and Trefethen [2], the computer scientists have the research challenge of creating high-level intelligent services that genuinely support e-Science applications. Such services, e.g., SGs, should go beyond straightforward access to computing resources, and also include support to construct and manage virtual organizations, as well as to manage the scientific data deluge in the scholarly cycle including hypothesis, experimentation, analysis, publication, research, and learning.

A large number of communities are therefore facing the challenge of building SGs. A recent collaboration resulted in the European Grid Infrastructure (EGI) Science Gateway Primer [3], where issues involved in SG design, implementation and operation are presented and discussed. According to this Primer, SGs are desktop or web-based interfaces to a set of applications and data collections. SGs comprise front-end and back-end components and offer services that facilitate access to computing and storage resources, as well as services provided by Distributed Computing Infrastructures (DCIs). Moreover, SGs support collaboration between researchers through exchange of ideas, tools and datasets. From the functional perspective, SGs are *SG frameworks* or *SG instances*. SG frameworks implement generic functionalities such as security, workflow and data management, and DCI access. Examples are WS-PGRADE/gUSE [4] and DARE [5] frameworks. SG instances are community-specific science gateways, with tailored interfaces and services for a specific application domain. SG instances can be be built using SG frameworks or with custom software stacks. See Section II for examples of SG instances.

The e-BioInfra Gateway [6] is a SG instance for large scale biomedical data analysis on the Dutch e-Science Grid [7]. It is designed to simplify usage of this infrastructure by biomedical researchers for the analysis of large datasets, and implemented based on a custom framework. It is deployed at the Academic Medical Center (AMC) of the University of Amsterdam (UvA), The Netherlands. It lowers barriers for users by providing services such as community Grid certificate and automatic file transport to and from the Grid resources. Since its deployment in production (early 2011), researchers have successfully performed large computations on the Dutch Grid infrastructure via the e-BioInfra Gateway with minor help from the support team. For example, Peters et al. [8], Wingen et al. [9], Rienstra et al. [10], and de Kwaasteniet et al. [11] have already published results of neuroscience research based on the data analysis performed via the e-BioInfra Gateway. The gateway structures the system information and allows for extensions with new data analysis methods. This enabled (external) developers to extend it with ten applications, six for medical imaging, three for next generation sequencing data analysis, and one for mass-spectrometry modelling. The e-BioInfra Gateway currently has 29 active users, and the largest usage so far is by the researchers from the Brain Imaging Center of the AMC (BIC [12]).

Although the current e-BioInfra Gateway can be considered a success story, our experience indicated the need for further

‡Corresponding authors: {s.shahand | s.d.olabarriaga}@amc.uva.nl

improvement respectively to the following aspects: *a*) support for data management; *b*) support for other types of DCIs, such as clusters and clouds; *c*) customizable interfaces to suit different user expertise, roles, and preferences; and *d*) sustainability of the adopted framework.

In this paper we discuss these experiences, which motivated the design of the next generation of the e-BioInfra Gateway. The new gateway design is generic (i.e., it is not specific to a particular research community), and it is based on the WS-PGRADE/gUSE SG framework [4], which facilitates access to heterogeneous DCIs. Additional features of the new gateway include data and information management, as well as support for meta-data that is used and generated during the execution of complex data processing. We describe the requirements, design and architecture of the new system, and discuss some initial results based on the prototype implementation for computational neuroscience, coined NeuSG. Although we focus on the neuroscience use case, the approach could be applicable to other domains as well.

II. Related Work

Design, development, and usage of SGs have gained interest and attention in the past few years. Several projects and initiatives have been started worldwide to develop SG frameworks and SG instances for diverse user communities [13]. For example, see the list of SGs on the websites of XSEDE (Extreme Science and Engineering Digital Environment) [14], EGI (European Grid Infrastructure) [15], and the SCI-BUS (SCIentific gateway Based User Support) [16] project.

The VIP (Virtual Imaging Platform) portal [17], the Charite Grid portal [18], and WeNMR gateways [19] are examples of SG instances based on custom frameworks. The MosGRID SG [20], VisIVO [21], and the Swiss Grid proteomics (iPortal) [22] portals are examples of SG instances based on SG frameworks (i.e., the WS-PGRADE/gUSE [4] in the case of these three). All of these SGs typically provide data and information management for a specific research community using custom solutions. Particularly in the field of medical imaging, two examples relate more closely to our work.

The data engine [23] of the CHAIN project [24] adopts the jSAGA implementation of the Simple API for Grid Applications (SAGA) standard to communicate with Grid resources for data storage. The related meta-data is stored in in-house databases. The CHAIN data engine is used in the CHAIN SG [25] and the DECIDE SG [26]. The DECIDE SG provides high-level services for computer-aided neurological diseases diagnosis and research on the European Research and Education Networks and the European Grid Infrastructure.

The neuGRID for you (N4U) Science Gateway [27] provides user-friendly access to N4U tools, algorithms, pipelines, visualization toolkits, and resources on various DCIs (Grid, Cloud, and Clusters) for medical imaging research, towards the cure of brain diseases, in particular Alzheimer's disease. The N4U Persistency Service registers distributed data from project partners into the N4U Information Base, which are then treated as a single data source.

In contrast to these SGs, our approach aims at generic services that are able to connect to existing data and information management services. These generic services are not dedicated to any domain-specific data type or format and try to remove the burden of moving files around from the user shoulders.

III. Background and Motivation for a new gateway

In a nutshell, the current e-BioInfra Gateway works as follows (see more details in [6]): the user authenticates with username and password, selects the application to run, selects the input files and other parameters, and starts a so called *experiment*. She/he can then monitor the experiment and, when finished, retrieve the results. The processing on grid resources is performed by the MOTEUR [28] workflow management system (WfMS), and provenance information is kept about the experiments. Because medical imaging data files are large, their transport is not done directly via the e-BioInfra Gateway web interface, but via an FTP directory that is located in the trusted network of the hospital. Therefore, for neuroscience applications the user uploads the data to the server before performing the steps above, and retrieves the results from the same place when the experiment is completed.

In these around two years of experience with gateway extension, operation, and user support, we faced challenges discussed below.

A large number of errors are caused by invalid input data. Users typically have difficulty to prepare files for processing with the gateway applications, which currently involves steps for file (re-)formatting, naming, transport, and also being aware of the data types that can be processed by each application. Although these problems are significantly reduced after training or reading the user manual, the data preparation and transport process should be improved with further automation.

Originally the e-BioInfra Gateway was meant to facilitate access grid resources. In the past years other resources have become available for research, such as local clusters at the AMC, a High-Performance Cloud, and GPU clusters. The current WfMS does not interface with clouds, so another solution is required to exploit these additional resources.

The current gateway supports two user profiles, end-users and administrators. We noticed, however, that additional profiles could be better supported with (combinations of) customized views of the various services [29]. For example, end-users can have different levels of expertise, or application support can be provided by members of the user community (and not necessarily only system administrators). Therefore a more flexible framework is needed to manage users, their roles and interaction, and viewing preferences.

Finally, we noticed the need for adopting a more sustainable software stack. Although our custom framework fulfilled the needs at first, as a small research group it is difficult to maintain and extend it. In particular, keeping up with all the developments related to DCIs requires significant effort and expertise that can be achieved by bundling forces across SG communities, such as done in the SCI-BUS project [16].

IV. Requirements Analysis

We described the typical phases of computational neuroscience studies in [30], which in summary include *study*

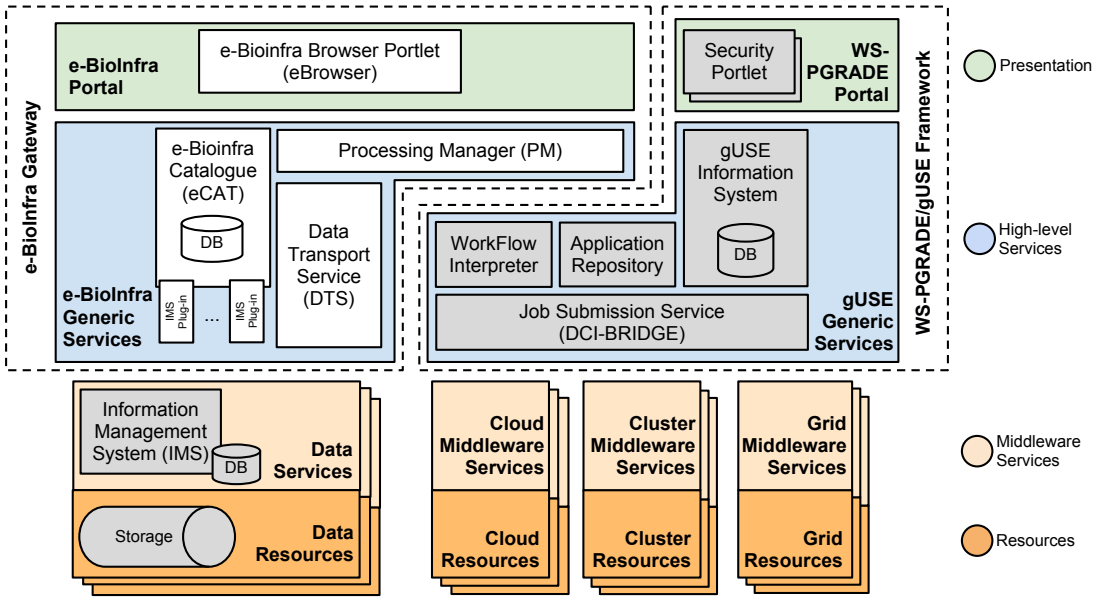

Fig. 1. Layered architecture of the e-BioInfra Gateway based on the WS-PGRADE/gUSE generic SG framework. Grey boxes represent existing third-party components, and white boxes denote components added for complementary functionality. See text for more details.

design, data acquisition, data handling, processing, analysis, and publication. Based on the analysis of these phases, the actors who are involved in each phase, and the tasks that they perform, in that paper we identified the properties and functionalities of SGs to support computational neuroscience research communities. In summary, the required properties and functionalities include: sharing of data and methodology; satisfying security and privacy regulations; scalable, transparent, and flexible management of storage and computing resources; literature discovery; collaboration support; meta-data, data, workflow, and provenance management; and visualization.

The current gateway [6] covers a subset of these requirements, namely: transparent authentication and authorization with Grid resources; flexible and efficient data transfer between local and Grid storage for files without user intervention; workflow processing management, including logging and monitoring; and an extensible set of applications for various biomedical domains. For the new gateway we focused on the following additional functionalities, in particular to further support *data handling*:

1) Unified, secure, and easy access to data and related meta-data stored on heterogeneous infrastructures and repositories. Users should be able to transparently query, explore, process, and analyse data from a single interface, without bothering about the data location or format, or how it is retrieved for further processing.
2) Automatic data format conversion and preprocessing according to pre-defined rules. For example, pseudonymisation and format conversion are automatically performed when new data is imported into the system.
3) Automatic and interoperable file transport and processing on different infrastructures (e.g., data servers, grid, cloud). Low level technical details are hidden from the users, such as different communication protocols, middleware services, and authorization mechanisms.
4) Automatic provenance information collection about the methods, parameters and input files used for processing.
5) Single sign-on facility to authenticate and authorize transparently to various computing and storage resources using user or community credentials.

In addition to these functionalities, we aimed for a gateway that is:

1) *extensible*, to easily accommodate new types of data or compute resources, applications, and user groups;
2) *customizable*, to be able to support different research communities and user profiles;
3) *scalable*, to gracefully support the growth of user community and its needs for resources, as well as infrastructures capacity and heterogeneity; and
4) *sustainable*, by using a community-driven SG framework.

V. System Design and Implementation

Figure 1 illustrates the layered architecture of the new e-BioInfra Gateway. At the bottom, the Resource layer (dark orange) with several DCI (e.g., local clusters, Grid and Cloud) and data resources (e.g., Radiology research data server). These resources are utilized through Middleware Services contained in the second layer (light orange). High-level Services contained in the third layer (blue) provide an abstraction to interact with the middleware, such as workflow management and data transport. Finally, the Presentation layer (green) contains the interfaces for user interaction. The two topmost layers (green, blue) are implemented using generic SG framework components provided by WS-PGRADE/gUSE (at the right), as well as a new data-centric SG framework that complements the functionality of WS-PGRADE/gUSE for the specific case of NeuSG (at the left).

Figure 2 illustrates the systems that host these components respectively and their network location. Due to security regu-

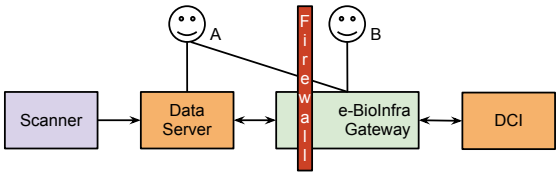

Fig. 2. Hosts and services of NeuSG and their network location: inside or outside the AMC firewall. The e-BioInfra Gateway is located in the demilitarized zone. User *A* is within the firewall boundaries and can access the data directly; user *B* is outside the firewall boundaries and therefore only has access to the meta-data and processing resources.

lations for processing medical research data, some services are hosted inside the hospital firewall. The data is generated by the scanner and directly imported into a Data Server located inside the firewall, which keeps both the raw data and the meta-data. The e-BioInfra Gateway is located in the demilitarized zone (DMZ) of the AMC network, which means that only some of its services are visible from outside the network. In Figure 2, both users *A* and *B* can browse meta-data, start and monitor data processing via the gateway, but only user *A* can download and view the medical imaging data. The raw data itself can only be accessed by the user directly from the Data Server, or by privileged services of the e-BioInfra Gateway.

Below we further detail the components that are more relevant for a data-centric SG, namely data services and the new components illustrated as white boxes in Figure 1. For completeness we briefly introduce the WS-PGRADE/gUSE SG framework, and finally, we describe the interaction between these components.

A. Data Services

Management of biomedical research data, with its growing size and complexity, requires domain-specific *Information Management Systems (IMSs)*. There are several IMSs that address challenges such as management of biomedical research data and meta-data, electronic data exchange, archival and security, and the research communities usually already adopt such systems routinely. Additionally, every community has its own procedure to implement rules and regulations regarding the protection of biomedical research data, as well as policies for data sharing and archiving. Therefore, instead of replicating such efforts, we decided to rely on existing, external, biomedical research data and meta-data resources, as well as on their own security mechanisms and policies. In this way, the research community itself provides and manages the IMS, defining data ownership, access policies, and regulating data confidentiality and privacy methods such as pseudonymisation. The IMS is connected to the e-BioInfra Gateway by agreement between the community and the gateway providers, and the data becomes available for processing at the gateway for authorized users only.

A popular IMS for medical imaging data and meta-data is the eXtensible Neuroimaging Archive Toolkit (XNAT) [31]. XNAT is an open source IMS that offers an integrated framework for storage, management, electronic exchange, and consumption of medical imaging data and its complementary meta-data. XNAT provides a rich communication layer based on a RESTful API of resource-oriented web services. Due to these qualities, XNAT has been deployed in the Radiology department of AMC and connected to the NeuSG as first supported IMS.

B. WS-PGRADE/gUSE SG Framework

WS-PGRADE/gUSE SG framework [4] is an open source, workflow- and service-oriented framework that facilitates development, execution, and monitoring of scientific workflows on DCIs. It comprises the WS-PGRADE portal, and the Grid User Support Environment (gUSE) services. WS-PGRADE is based on the Liferay portal framework, which provides rich facilities for community management and customizable user interfaces. gUSE provides high-level services to access various DCI resources. These qualities motivated the choice for this SG framework to implement our gateway.

The most relevant gUSE services for our gateway are:

- *Job submission service or DCI-BRIDGE:*[1] provides flexible and versatile access to a large variety of DCIs such as grids, desktop grids, clusters, clouds and service-based computational resources. It also handles authentication and authorization to the configured DCIs transparently.
- *Workflow Interpreter:* parses workflows, submits jobs to the DCI-BRIDGE, and retrieves their status for monitoring and fault-tolerance.
- *Application Repository:* stores ready-to-use tested and configured workflows. These workflows are exported to the application repository by workflow developers, from where they are imported into user space for execution.
- *gUSE Information System:* stores configurations of gUSE services and workflow related information such as workflow executions and their jobs status.

The WS-PGRADE/gUSE framework also provides two Application Programming Interfaces (APIs) to create SG instances. We used the *Application Specific Module (ASM)* API to utilize gUSE services, more specifically the Application Repository, to manage and share workflows among users, and the Workflow Interpreter, to submit workflows.

The WS-PGRADE portal also offers a set of generic portlets to interact with gUSE services via web-based graphical user interfaces. For example, users can manage their credentials, which are required to authenticate and authorize to DCIs, via security portlets. See [4] for the complete description of WS-PGRADE/gUSE services and portlets.

Currently the WS-PGRADE/gUSE framework does not have any facility to connect to external IMS resources. Moreover, its current data management facilities are also limited. The data-centric e-BioInfra Gateway tries to bridge this gap with additional components described below.

C. e-BioInfra Gateway data centric framework

The core of the new e-BioInfra Gateway is made of the following components: e-BioInfra Catalogue (eCAT), Data Transport Service (DTS), Processing Manager (PM), and e-BioInfra

[1] According to [4], the DCI-BRIDGE has been moved out of the gUSE layer to highlight that it is directly accessible via the standard OGF BES job submission interface. Here we utilize a different conceptual framework to illustrate the architectural layers of the system, thus we consider it as part of the gUSE generic services.

Browser Portlet (eBrowser). They are loosely coupled and communicate via well-defined APIs, an approach that paves the road towards a service-oriented architecture and facilitates their reuse in other gateways. These components are deployed in the same environment alongside WS-PGRADE/gUSE components and work together to implement the NeuSG functionalities.

D. e-BioInfra Catalogue (eCAT)

The eCAT has been designed to facilitate the data management functionalities at the gateway. It is a central information store for user-specific configurations such as IMS hosts and the user's credentials to access them. eCAT defines and implements a data model to manage system-level information, with the following main entities: `User`, `Project`, `Data`, `Application`, `Processing`, and `View preferences` (see Figure 3 for their relationships).

eCAT provides an aggregated and user-specific view of `Data` entities that each user has access to a given IMSs. Note that eCAT is not meant to duplicate meta-data that is already stored on IMSs; instead, it only stores pointers to such information on IMSs. It retrieves and stores meta-data on IMSs through the respective *IMS Plug-ins*, which are software modules attached to eCAT to enable programmatic communications with a specific IMS. The only exceptions are some meta-data that are specific to user activities on the gateway, which are not possible, nor of direct interest of research communities, to store in their IMSs. For example, location of data replicas on a DCI and user `View preferences` are such meta-data that are only stored in the eCAT database.

`Data` entities are included in, and processed within, the scope of `Project` entities. When possible, `Projects` are also in sync with those on IMSs. Each `User` has access to some `Applications`, which are tested and ready-to-use workflows. When a `User` processes a certain `Data` with a specific `Application`, the information about this activity is captured by eCAT as a `Processing` entity. The provenance information about the `Data` consumed and produced during a `Processing`, the parameters, and the latest status of processing, are also stored in the eCAT database. eCAT also provides necessary information to transport the results produced by a data processing to the respective IMS, if possible together with the provenance information. eCAT is accessed by other system components (PM, DTS, and eBrowser) through its API.

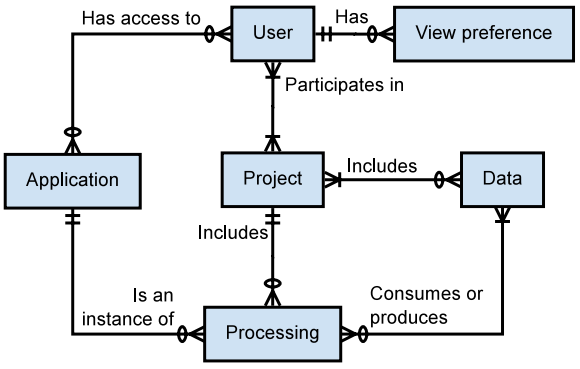

Fig. 3. Simplified entity-relationship model of the information stored in the e-BioInfra Catalogue.

E. Processing Manager (PM)

The PM takes care of submission and monitoring of data processing applications, which are defined as workflows that are executed by the gUSE Workflow Interpreter. The PM instructs to the DTS to transport input files from the IMSs to the storage resources of the DCI on which the processing is performed, and to transport the results back to the IMS. The PM imports the workflow from the gUSE Application Repository and configures it with the physical location of input data before submission.

F. Data Transport Service (DTS)

The DTS transports data between IMSs and storage resources on DCIs. This service contacts the eCAT to determine how to authenticate the IMS on behalf of the user, how to authenticate with the storage resources of the DCI (possibly with community credentials), and how to access data on both. It autonomously performs the data transfer using third-party mechanisms as much as possible to avoid bottlenecks. If some data has been replicated on a DCI, the location of that replica is stored in the eCAT and retrieved later.

G. e-BioInfra Browser Portlet (eBrowser)

Unlike the previous components, the eBrowsert is part of the presentation layer. It provides a web-based user interface to interact with all the e-BioInfra generic services. Instead of contacting the services directly, eBrowser retrieves information from eCAT to provide a unified view to scientists to browse data, projects and data processing instances. eBrowser essentially enables scientists to start, manage, and monitor data processing (through PM), as well as to configure viewing and interaction preferences with the gateway.

H. Component Interactions

Figure 4 illustrates the interactions between users and the e-BioInfra Gateway, as well as the interactions between underlying components. User actions are expressed via the eBrowser and trigger interactions between other high-level components (i.e., PM, DTS and eCAT) and lower-level components (i.e., gUSE and XNAT IMS). Details of these interactions are presented below.

Upon successful authentication with the WS-PGRADE portal, the user gets access to the eBrowser portlet. New users need to configure an IMS endpoint by providing the URL of the IMS, its type (e.g., XNAT), and recording their username and password securely. These configurations are collected by the eBrowser and sent to eCAT for validation and storage. After this configuration step, the following takes place when the user logs into the e-BioInfra Gateway

1) At first the user sees a list of her projects. To display this list, eBrowser sends a request to eCAT, which authenticates on behalf of the user to all registered IMSs and generates a unified list of all projects that are accessible by that particular user.
2) Similarly, when the user selects a project, the eBrowser sends a request to eCAT, which queries meta-data on the IMS to produce the list of all data entries in that project.

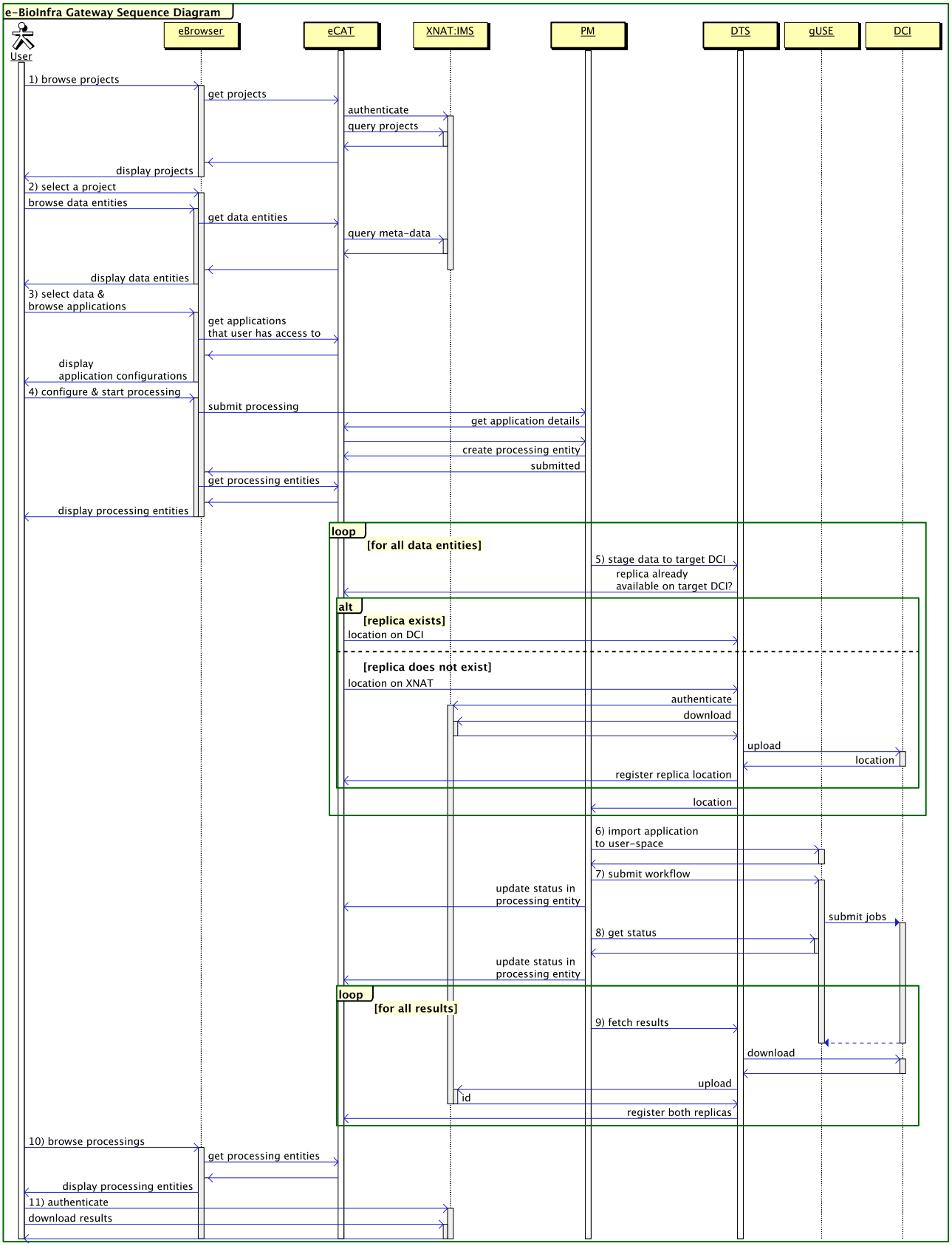

Fig. 4. Sequence diagram illustrating the interactions between the user and the various NeuSG components. After authentication, the user can browse projects, select data based on meta-data, select an application to run on these data and start new processing, monitor processing, and download results.

3) The user then selects data entities that she wishes to process, and browses for available applications. The eBrowser retrieves and displays the list of applications accessible to the user. The user selects an application and the eBrowser displays configurations for that application (e.g., application parameters).
4) The user configures the application and starts a new data processing. The eBrowser collects the provided configuration and submits a processing request to the PM. The PM consults eCAT to find the details of the selected application, namely the DCI to run it and the arguments that need to be configured for its execution (e.g., input files and parameters). The PM validates and creates the processing entity in eCAT, from which the eBrowser can later retrieve and display to the user for browsing, management, and monitoring purpose.
5) The PM further instructs the DTS to move the required input data to the target DCI. The DTS contacts eCAT to determine if those data already have a replica on the target DCI. If no replica is available, the eCAT provides DTS with the IMS endpoint configurations (including authentication token) and location where it can retrieve the input data. The DTS then uses this information to authenticate on behalf of the user to the IMS and download the input data. Similarly, it retrieves user authentication tokens for the target DCI to upload input data (not shown in the diagram). Finally the DTS registers in eCAT the location of the file replica in the DCI and returns it to the PM.
6) After all data have been staged to the target DCI, the PM imports the application from gUSE via the ASM API into the user-space, and configures it with the physical location of input data and user-specified parameters.
7) Having everything in place, the PM starts the data processing by submitting the configured application (workflow) to gUSE via the ASM API, and updates the processing status in eCAT. The gUSE Workflow Interpreter parses the workflow, generates corresponding jobs, and submits them to DCI-BRIDGE. The DCI-BRIDGE retrieves user authentication tokens for the target DCI to submit jobs on behalf of the user to the target DCI.
8) The PM periodically updates the information in eCAT based on the status reports from gUSE, which is then reflected in the interface of the eBrowser for monitoring.
9) Typically, each processing consists of multiple data to be processed. When the processing of some data are finished, their results are immediately stored in the specific IMS via the DTS. Thereby the user can check results even before the entire processing is complete.
10) The user browses, manages, and monitors the processing via the eBrowser. eBrowser contacts eCAT to get information about processing entities, including status.
11) The user is forwarded to the IMS directly to access and download processing results via a link from at the gateway interface.

VI. Discussion

In the new generation of the e-BioInfra Gateway we tried to bridge the gap between scientists, data services, and DCIs. We aimed for a data-centric gateway in which everything is organized around "data". Now scientists can use the gateway not only to browse their data, which can be potentially stored on several IMSs and described by rich meta-data, but also to perform large scale data processing on DCIs. This can be done without getting involved into low-level details, such as transporting files, as it was the case in the previous generation.

The previous generation was built based on the Spring framework, it only supported the Dutch Grid infrastructure, and it lacked facilities for user interface customization or community support. In contrast, the new generation of the e-BioInfra Gateway is built based on the WS-PGRADE/gUSE SG framework, which itself is built on the Liferay portal framework. Liferay provides facilities for user management, community management, and community support (e.g., online forum). Moreover, it also facilitates the construction of customizable web-based user interfaces that are required to suit needs of each user (community) based on their profile, expertise, and roles. The WS-PGRADE/gUSE SG framework provides high-level generic services to manage workflows, enact them to various DCIs, and monitor their execution. These services allow for functional scalability and interoperability between various DCIs. Additionally, the WS-PGRADE/gUSE framework is an actively maintained and developed open-source project, which allows the development team of the e-BioInfra Gateway to concentrate on its community-specific features, and makes the gateway operation more sustainable.

Currently only XNAT is supported as IMS. Several other data management platform alternatives meet the research requirements, although XNAT is of special interest due to its support for medical imaging, and its adoption by the AMC neuroscience research community. It has been particularly designed for managing standard medical imaging data as the core of its functionalities. In addition, its archiving and integrating capabilities, data model flexibility, ease of use and the highly active community of users/developers makes it a relevant asset. Note however that the new e-BioInfra Gateway has been designed to support multiple and heterogeneous IMSs, and it is not dependent on XNAT.

The eCAT contains much meta-data about the system level (viewing and processing), but it is completely dependent on an external IMS for the data. If the IMS is not available, the user cannot perform any data-related activity, such as browsing or selecting files. We have considered duplicating the meta-data on the eCAT, both for efficiency and fault-tolerance reasons, but we concluded that the synchronization of the two systems would be too time consuming. Moreover, we chose to keep the access control to the Data Server completely in the hands of the community administrators, which, due to the required expertise, can be different persons than the SG administrators. This helped us build trust between the systems, which is a known critical factor to connect such systems to open infrastructures such as grids and clouds.

We used WS-PGRADE/gUSE as SG framework, which in principle provides the workflow management and portal functionalities needed for the NeuSG. After a learning phase, during which the concepts of the framework were better understood by the team, we observed that the usage model of the framework differs from our needs in some cases, which has led us to develop our own *processing manager* component. This has the goal of translating high-level "data processing" commands into low-level data transports, which are performed by the *data transport service*, and calls to the gUSE ASM

API. At first this introduces overhead, but at the same time it provides sufficient isolation from aspects regarding this particular WfMS, and allows us to consider other WfMSs in the future.

The development of eBrowser viewing portlets was also simplified by the decision to have all user interaction to take place using information available on the eCAT. This approach requires all software components to register all activity on the eCAT, but it decouples the viewer from all the other components accordingly. This reduces dependencies between the system components and simplifies its implementation and maintenance. Moreover, it make the eCAT as a natural provenance data repository for the activity carried out at the gateway.

VII. Conclusion and Future Work

The implementation is being completed, and the new gateway will be released soon (April) for evaluation by AMC BIC users. The portfolio of applications will be enriched (currently there are only two), and the eBrowser will be extended (currently only basic browsing functionality is available). At a second step, the gateway will be disseminated in training events, and become open to the whole neuroscience community of the University of Amsterdam. This step will require inclusion of other IMSs, for example other XNAT instances or even other systems, as well as extending the eCAT with federated services for accessing (and/or querying) multiple IMSs. Increasing number of users and data will require further development of instruments for strong community support, communication and access control tools, part of which are supported by Liferay. Moreover, semantic content annotation (ontologies), as well as adding knowledge and integrating it with existing data, could enable further automation of the data processing to reduce even more human intervention in the analysis of large quantities of biomedical data.

Finally, we kept bioinformatics researchers in the loop during the requirement analysis, design, and implementation of the gateway to assure that the resulting SG is generic enough to support bioinformatics research community with minimal additional effort. Although in this paper we are focused on the computational neuroscience applications, the same concept and software components are being used to develop a SG for analysis of DNA sequencing data.

Acknowledgment

This work is financially supported by the COMMIT project "e-Biobanking with imaging for healthcare" funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organisation for Scientific Research, NWO), the SCI-BUS project funded by European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 28348, and the HPCN UvA project "Computational Neuroscience Gateway" funded by the University of Amsterdam.

References

[1] D. De Roure *et al.*, "The semantic grid: Past, present, and future," *Proceedings of the IEEE*, vol. 93, no. 3, pp. 669 –681, march 2005.

[2] T. Hey and A. E. Trefethen, "Cyberinfrastructure for e-science," *Science*, vol. 308, no. 5723, pp. 817–821, 2005.

[3] E. G. I. Science Gateway Virtual Team, *Science Gateway Primer*. EGI (European Grid Infrastructure), 2012.

[4] P. Kacsuk *et al.*, "WS-PGRADE/gUSE Generic DCI Gateway Framework for a Large Variety of User Communities," *Journal of Grid Computing*, vol. 10, no. 4, pp. 601–630, 2012.

[5] S. Maddineni *et al.*, "Distributed Application Runtime Environment (DARE): A Standards-based Middleware Framework for Science-Gateways," *Journal of Grid Computing*, vol. 10, no. 4, pp. 647–664, 2012.

[6] S. Shahand *et al.*, "A grid-enabled gateway for biomedical data analysis," *Journal of Grid Computing*, vol. 10, no. 4, pp. 725–742, 2012.

[7] "The BiG Grid Project website," http://www.biggrid.nl.

[8] B. D. Peters *et al.*, "Polyunsaturated fatty acid concentration predicts myelin integrity in early-phase psychosis," *Schizophrenia Bulletin*, 2012.

[9] G. A. van Wingen *et al.*, "Persistent and reversible consequences of combat stress on the mesofrontal circuit and cognition," *Proceedings of the National Academy of Sciences*, vol. 109, no. 38, pp. 15 508–15 513, 2012.

[10] A. Rienstra *et al.*, "Symptom validity testing in memory clinics: Hippocampal-memory associations and relevance for diagnosing mild cognitive impairment," *Journal of Clinical and Experimental Neuropsychology*, 2012.

[11] B. de Kwaasteniet *et al.*, "Relation between structural and functional connectivity in major depressivedisorder," *Biological Psychiatry*, no. 0, pp. –, 2013.

[12] "The BIC (Brain Imaging Center) at the AMC (Academic Medical Center) website," http://www.lebic-amc.nl.

[13] T. Kiss, "Science gateways for the broader take-up of distributed computing infrastructures," *Journal of Grid Computing*, vol. 10, pp. 599–600, 2012.

[14] "The XSEDE (Extreme Science and Engineering Digital Environment) website," http://www.xsede.org.

[15] "EGI (European Grid Infrastructure) Science Gateways," http://www.egi.eu/services/support/science-gateways/index.html.

[16] "The SCI-BUS (SCIentific gateway Based User Support) Project website," http://www.sci-bus.eu.

[17] T. Glatard *et al.*, "A virtual imaging platform for multi-modality medical image simulation," *Medical Imaging, IEEE Transactions on*, vol. 32, no. 1, pp. 110 –118, jan. 2013.

[18] J. Wu *et al.*, "The charité grid portal: User-friendly and secure access to grid-based resources and services," *Journal of Grid Computing*, vol. 10, pp. 709–724, 2012.

[19] T. Wassenaar *et al.*, "WeNMR: Structural Biology on the Grid," *Journal of Grid Computing*, vol. 10, pp. 743–767, 2012.

[20] S. Gesing *et al.*, "A single sign-on infrastructure for science gateways on a use case for structural bioinformatics," *Journal of Grid Computing*, vol. 10, pp. 769–790, 2012.

[21] E. Sciacca *et al.*, "VisIVO Workflow-Oriented Science Gateway for Astrophysical Visualization," in *Proceedings of the 21st Euromicro International Conference on Parallel Distributed and Network-Based Processing*, 2013.

[22] P. Kunszt *et al.*, "The swiss grid proteomics portal," in *Proceedings of the Second International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering*, 2011.

[23] M. Fargetta *et al.*, "A data engine for grid science gateways enabling easy transfer and data sharing," Presentation in the EGI community Forum 2012, March 2012.

[24] "The CHAIN (Co-ordination and Harmonisation of Advanced e-INfrastrucures for Research and Education Data Sharing) Project website," http://www.chain-project.eu.

[25] "The CHAIN Science Gateway," http://science-gateway.chain-project.eu.

[26] V. Ardizzone *et al.*, "The decide science gateway," *Journal of Grid Computing*, vol. 10, no. 4, pp. 689–707, 2012.

[27] "The N4U (neuGRID for you) Project website," http://neugrid4you.eu.

[28] T. Glatard *et al.*, "Flexible and Efficient Workflow Deployment of Data-Intensive Applications On Grids With MOTEUR," *International Journal of High Performance Computing Applications*, vol. 22, no. 3, pp. 347–360, Aug. 2008.

[29] S. Shahand *et al.*, "Front-ends to Biomedical Data Analysis on Grids," in *Proceedings of HealthGrid 2011*, Bristol, UK, 2011.

[30] S. Shahand *et al.*, "Integrated Support for Neuroscience Research: from Study Design to Publication," in *Proceedings of HealthGrid 2012*, Amsterdam, NL, May 2012.

[31] D. Marcus *et al.*, "The extensible neuroimaging archive toolkit," *Neuroinformatics*, vol. 5, pp. 11–33, 2007.