The CARMEN Neuroscience Server

Paul Watson[1], Tom Jackson[2], Georgios Pitsilis[1], Frank Gibson[1], Jim Austin[2], Martyn Fletcher[2], Bojian Liang[2], Phillip Lord[1]

[1]School of Computing Science, Newcastle University, UK
[2]Department of Computer Science, University of York, UK

Abstract

Understanding the brain is one of the major scientific challenges. It requires the capability to synthesize a detailed and applicable understanding of the way in which information is encoded, accessed, analysed, archived and decoded in neuronal networks. Data is difficult and expensive to produce, but is rarely shared and collaboratively exploited. The main reason for this is that a proliferation of techniques produce voluminous data in a variety of heterogenous and proprietary formats; this is then locally described and curated and is often not computationally amenable. The EPSRC CARMEN e-Science Pilot project (www.carmen.org.uk) is addressing these challenges by leveraging e-Science infrastructure and expertise to support the virtual integration of research teams, and multi-modal experimentation. CARMEN will allow data sharing and integration, supported by metadata and an expandable range of services accessible to users for raw, transformed and live experimental data. Achieving this requires progress in a number of areas including: the ability to store, curate and deploy services as well as data; standardised metadata for neuroscience; and, advanced tools for searching and visualising time-series and related data. This paper gives an overview of the CARMEN infrastructure, and illustrates its functionality by describing the application of an early prototype to a specific neuroscience scenario.

1. Introduction

Understanding the brain is one of the major scientific challenges. Tackling it requires the development of a detailed and applicable understanding of the way in which information is encoded, accessed, analysed, archived and decoded in neuronal networks. This is not only fundamental to the objectives of computational and experimental neuroscience, but has major application to computer science (*neuromorphic* and *neuromimetic* systems), nanotechnology (neuronal interfaces, neuroprostheses, and biosensors), electronic engineering and informatics (data handling and design strategies for multi-sensor systems), and pharmacology (*in silico* drug development).

Information flow in neuronal networks is made up of a series of component processes which can be investigated using a range of electrophysiological and imaging techniques. Data is difficult and expensive to produce, but rarely shared and collaboratively exploited. A proliferation of techniques produce data that is voluminous, proprietary, locally described and curated. It is therefore difficult to integrate and often not computationally amenable. It is also uncommon for the recipient data analyst to apply analysis algorithms to more than one neuronal system. As a result, there is:

(a) a shortfall in analysis techniques that can be applied robustly and effectively across neuronal systems;
(b) an absence of readily accessible data to support development of analysis techniques;
(c) no structured curation and optimisation of data;
(d) only sporadic interaction between disparate research centres with complementary expertise (e.g. data collection vs. data analysis);
(e) very limited understanding of the informatics solutions that could be applied to broaden the science by integrating data across spatiotemporal scales.

The challenge is thus to instigate a step change in the research methodology. Recent developments in e-Science make it timely to address this challenge, and that is the aim of CARMEN, an EPSRC e-Science Pilot project running from 2006-10. This will be achieved through the provision of an e-Science infrastructure to support cross-modal data sharing and integration, supported by metadata and an expandable range of services accessible to users for raw, transformed and live experimental data. These innovations will create a virtual neuroscience laboratory that ties together experimental and computational neuroscientists.

In this paper we present the architecture of the CARMEN CAIRN – the core infrastructure being designed and deployed to store and process information produced by neuroscientists

(Section 2). We then outline one of the experimental scenarios that are being used to drive the design (Section 3) before describing how this is supported by an early prototype of the CAIRN (Section 4). Finally we draw conclusions from this early work.

2. Architecture

An e-Science infrastructure is being built to fulfill the neuroinformatics requirements. At its heart is a CARMEN Active Information Repository Node (CAIRN), whose role it is to store and process the data (Figure 1). Its design builds on existing e-science components produced or utilized in earlier projects in which the project partners were involved (especially, myGrid [1], DAME [2], BROADEN [3] and Gold [4]). The rest of this section describes the architecture and components of the system.

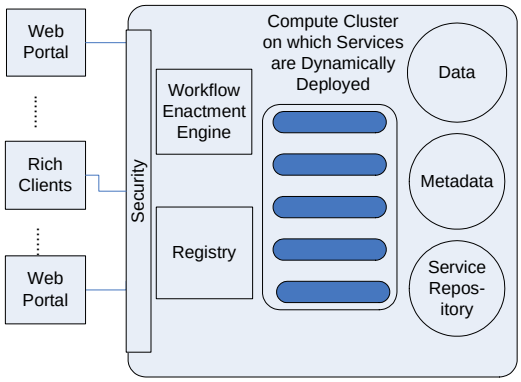

Figure 1. The CARMEN CAIRN (CARMEN Active Information Repository)

2.1 Storage and Management of Data

The repository will hold both raw voltage signal data (e.g. collected by Multi-Electrode Array recording) and image data (e.g. resulting from the internal chemical processes of the neuron and the activity dynamics of large neuronal networks). Services will be provided to import and export data in a variety of proprietary (raw) data formats (current neurophysiology instrumentation does not produce a uniform data format and is, in many cases, manufactured by small, autonomous companies) and convert them into a common format that supports analysis.

Due to the large volume of data that will be produced by the neuroscience experiments, there is an initial requirement to hold in the region of 50TB of data. Whilst the primary data will be held in file storage in a Storage Resource Broker (SRB [5]), the derived data will be stored in a database in order to allow researchers to exploit the powerful functionality (especially indexing and querying).

2.2 Metadata.

Metadata is critical to ensure that the stored data can be discovered and understood. To allow CARMEN to provide rich metadata, it is first essential to ensure that enough information can be made available to provide an understanding of the experimental context in which the data was gathered. Therefore, the tools provided for importing data into the repository will enable the user to specify and reuse descriptions of the experimental context and conditions. Secondly, it is equally important that data derived by analysis can be accessed in the same way. Therefore, the analysis services provided by CARMEN will describe themselves in terms of their domain functionality. During workflow enactment this metadata will be used to generate automatic provenance traces.

We will build on previous work from the myGrid project, and use LSIDs (Life Science Identifiers – [6]) to draw relationships between data and metadata.

Support for service metadata will be based on previous work on ‘Feta’ in the myGrid project. The Feta data model will be modified to allow it to describe neuroinformatics services. This will allow the OMII’s Grimoires [7] registry to be used for storing and searching over both data and service descriptions. We will exploit the linkage between data and service descriptions to enable users to intelligently discover appropriate analysis services for both primary and derived data.

Both of these core components of the metadata will be used to record data provenance. We will extend the workflow enactment engine to automatically gather and store information about data derivation. Where possible, we will augment this with knowledge of the user’s experimental context, ensuring that the purpose of the analysis is also stored. When combined with the basic experimental metadata, this will provide a very rich data discovery environment.

2.3 Data Analysis Services

The CAIRN will be a repository for the long-term storage and curation of analysis services as well as data. Typical services will be for spike identification, statistical analysis and visualization. These are packaged as WS-I conformant Web Services (http://www.ws-i.org/) so there is a common way of communicating with them. Scientists upload

their services in a deployable form into the CAIRN where they stored, and metadata about them is entered into a service registry. This ensures that services are preserved so that computations can be re-run, and services re-used, in the future (experiences in earlier e-science projects showed the danger of relying on the future availability of externally managed services). This approach of moving the computation to the data avoids the need to export the required data out of the CAIRN to a client in for processing (something that will be very, perhaps prohibitively, expensive when large datasets are involved).

The Dynasoar [8] dynamic service deployment infrastructure is used to deploy the services on demand from the repository onto the available compute resources when they are invoked. To meet changing loading requirements, services can be deployed on multiple nodes and requests load-balanced across the set of deployments.

Due to the high computational cost of running many of these services on large quantities of data (particularly in the guise of complex, parameter constrained models), access to large compute resources is required. Therefore, each CAIRN will contain a local compute cluster. Regional and National Grid resources, including the White Rose Grid and the Newcastle Grid will also be harnessed to provide additional resources when local capacity is exhausted.

2.4 Workflow Enactment Engine

Workflows provide a way to represent and enact processes, such as complex data analyses, that involve data being processed by multiple services. The workflow enactment engine is made available as a service in the CAIRN to allow external clients to submit workflows to be enacted close to the data on which they will operate. This removes the costly need to transfer data in and out of the CAIRN to an external enactor. The prototype described in the next sections utilises the OMII myGrid Taverna/Freefluo {Hull, 2006 #17} system.

2.5 Services for Annotation and Pattern Search in Time Series Data

Allowing users to efficiently locate patterns in the data is a crucial and novel requirement for the neuroscience that will be supported by CARMEN. Experiments will produce primary data that will be loaded into the CAIRN. Features form the data can then be identified, such as waves of electrical activity from a firing neuron, known as an action potential or “spike”.

These data may then be reprocessed to find groups of spike events that represent some correlated, associated event. Thus, the repository will hold time series data at a number of levels of abstraction. These data will only make sense if the events they contain (e.g. spikes) can be identified and annotated with ‘what, where and when’ metadata. Given the quantity and complexity of the data, neuroscientists will require tools to support this annotation process. For this, the project will build on the pattern search and matching technology developed by the DAME project. The Signal Data Explorer (SDE) [2, 10] can compare events in time series based on: existing events, those stored in a file, or new events drawn by the user. It has the ability to match across multiple time series (using a task planner tool) as well as allowing the user to tag the data with known events. To date, the main application of SDE has been searching vibration and performance data from aero engines. CARMEN will adapt this technology to detect events taking place within and across time series data within given timing conditions.

2.7 Security Infrastructure

The neuroscience community has a requirement to be able to control access to data and services. Without this, scientists would be reluctant to upload their data and services into the CAIRN. For example, researchers may wish their data to be accessible only to themselves and their collaborators until the point at which they have completed and published their analyses. Further, there is a requirement to support commercial organisations that will have a legal requirement to control, secure and protect access to the data they store. The security infrastructure will, therefore, deal with the authentication and authorization of users, and will be configured so as to enforce a mutual IPR agreement defined by the user community. For this we will utilise the Gold [4] security infrastructure, which uses XACML [11] assertions for role and task based access control. To remove the need for users to manipulate XACML directly, data owners will specify their security preferences through a graphical interface.

2.8 Client Access

A portal is being provided to allow users to interact with the CAIRN (e.g. to upload data and annotate it with metadata; to locate and

browse data; to create, run and monitor workflows etc.).

2.9 CAIRN Integration

It would be unrealistic to expect the global neuroscience community to use one CAIRN for all of their data. Instead, we anticipate that different communities, separated both by geography and focus, will create their own CAIRNs. CARMEN will, therefore, develop an integration layer that can present the data and services of a set of CAIRNs as if they were a single entity. This will be achieved by utilising and extending existing integration technologies, including Distributed Query Processing software from the OGSA-DQP [12] project. The SDE system already supports transparent searching across distributed data sets, though the project will enhance it to improve the intelligence of the search process.

3. Example Neuroscience Scenario

In this section we describe a neuroscience electrophysiology experiment that is acting as a use-case for CARMEN. We present an outline of the science, and describe how the experiment has been conducted to date. Next, in Section 4 we show how it is implemented using the CARMEN prototype. This has allowed the project to investigate how e-science technologies can be married together to facilitate data storage and analysis in neuroscience.

The objective of this experiment is to investigate the neural activity patterns in neonatal retina of the Cone Rod homoboX knockout mouse [13]**.** The ultimate goal is to use the CRX adult retina to develop a retinal implant consisting of a high density microelectrode array (MEA) to emulate vision by directly stimulating the ganglion cells. The output of this experiment is a visual representation of the neural activity on the retina.

3.1 Background

There are two types of neural activity that are recorded from the ganglion cells using a MEA. The first is spontaneous activity which disappears as the mouse develops with age, and the second is the intrinsic responses of a subset of ganglion cells to blue light. Our scenario focuses on the latter.

The working hypothesis of this study is that there is both spontaneous activity and intrinsic photosensitivity that the CRX retina expresses to compensate for the lack of visual experience through the visual photoreceptor pathway.

In this scenario we focus only on spikes generated by individual neurons. These usually have amplitudes of between 50-200 µV and do not exceed 5 ms in duration. As each electrode records signals from several cells simultaneously a spike sorting technique is used to relate spikes to individual neurons.

3.2 The Experiment

In a standard experiment recordings are taken using a Multi Electrode Array with 59 platinum disk electrodes. A Multichannel Systems MEA60 system [14] was used to performing recordings at a sampling rate of 25kHz. The recordings are amplified, digitized and stored in a PC attached to the MEA equipment that is running the MC Rack software (from Multichannel Systems). In this scenario, the response of the retina to light stimulation over a period of 120s was measured.

Later, off-line analysis of the data was performed starting with root-mean-square (RMS) noise removal (by applying a threshold value based on visual inspection) followed by spike detection using the crossing of the threshold value as the criteria for identification. At present this process is carried out using the software package MC Rack.

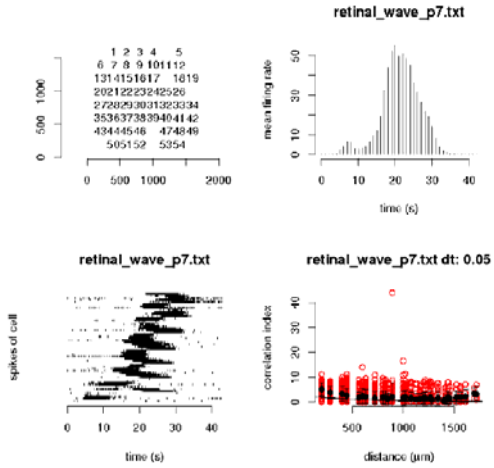

Figure 2. The Graphical Output of the Experiment. The four plots shown are from left to right, top to bottom: Geometry of the the array, Mean firing rate through the experiment, Spike trains of all channels and Correlation index [17].

3.3 Statistical Analysis

This step comprises the statistical analysis of the spike times file using the SJEMEA package

which is based on the R statistical programming environment [15].

SJEMEA analyses the spike time input files from *MC Data* and produces 4 graphs which present: (Figure 2)

1. (top left) The geometry of the array used in the MEA equipment showing the location of each active channel in the rectangular array in relation to the x and y axis.
2. (top right) The mean firing rate throughout the experiment, represented as the average number of spikes detected per channel.
3. (bottom left) The spike trains for each channel - each spike is denoted by a small vertical bar.
4. (bottom right) The correlation index plot, showing the correlation between the normalized number of synchronous spike pairs and the distance between every pair of active channels. Movies showing the spiking activity over time can also be created by users using the SJEMEA package.

4. Implementing the Scenario in the CARMEN Prototype

An advantage of basing the CARMEN infrastructure on the output of existing projects is that it was possible to quickly create a neuroscience-relevant prototype to explore options. In this section we describe how the above experimental analysis is now being performed using this prototype CARMEN infrastructure.

Data visualization and search capability are important elements of the CARMEN system ad these have been rapidly built into the system by deploying the Signal Data Explorer technology. The SDE [10] provides data visualisation, transformations and real-time search capabilities for complex signal data. The SDE was originally developed by the University of York and Cybula during the DAME [2] and BROADEN [3] projects. It provides the capability to search for patterns in temporal data signals across distributed repositories.

The interactive search capability is provided by the SDE via the distributed Grid data repositories which use:

- Distributed data management - the SDE interfaces directly to the Storage Resource Broker (SRB) from San Diego Supercomputer Center (SDSC).
- The pattern matching architecture consisting of Pattern Match Controller (PMC) and Pattern Match Engine (PME) components.

The raw data to be stored by the CARMEN CAIRN nodes consists of spatiotemporal signals expressed, either as time-series recordings collected by single electrodes or MEAs, or as image files, collected at regular intervals using various optical recording techniques. An example of the visualization capability is shown in Figure 3. SDE can support viewing of multi-channel data simultaneously, and in a highly interactive manner. This is a crucial feature for off-line analysis of the complex experimental data and for real-time computational steering of experiments.

A windowing capability permits auxiliary data views to be opened which permit the user to zoom into or out of data, providing macro and micro views, as well as allowing the user to “play and zoom” very large data sets. The first iteration of the CARMEN SDE also provides the capability to apply a range of data filters, provide various viewing modes to complement the main view and identify spike times from the raw data. In the longer term, the SDE will integrate with other spike detection and sorting algorithms being produced within the other work packages.

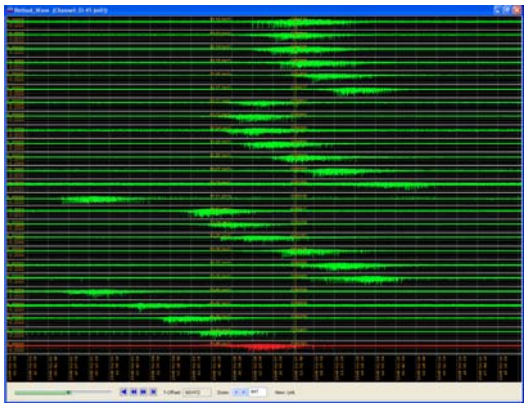

Figure 3. The Signal Data Explorer displaying signals from multiple electrodes

In addition to the visualization functions the SDE provides an interactive and intuitive search capability, such that features of interest can be located in archived and local data sets. The search process is feature driven, in that the user can highlight a region of interest in a time series signal and request a pattern matching process to be carried out against the target data sets. Similarity measures are used to provide a

ranking system that can score results for the search process. The architecture that underpins the search process has already been proven within the context of the DAME system and has been shown to be scalable to terabyte data sets. An example of a search process across an archive of multi-channel electrode sensor data is shown in figure 4.

The SDE can currently search local and distributed data stores, and interfaces directly onto the data archives that are being held on the CAIRNs. The raw data and the spike timing data is uploaded by the experimenters into the CARMEN CAIRN where it is stored and catalogued within the SRB. This primary data (and later, the results of the analysis) are uniquely identified by URIs (it is expected that later prototypes will use the LSID form of URIs for naming).

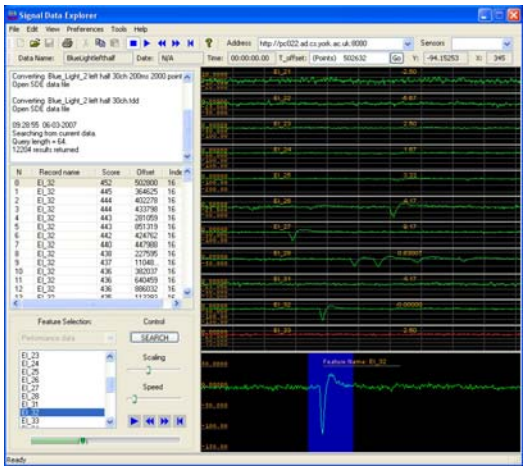

Figure 4: The pattern matching search process carried out within the SDE tool on MCD data.

The remaining processing of the experimental data is performed by a workflow that calls the analysis code wrapped as a Web Service. This particular scenario was expressed as an OMII Taverna workflow [9] as in Figure 5. As is the case in this example, due to the large quantities of data to be processed, we expect that services will input and output the URIs of data, rather than the data itself (e.g. pass by reference rather than by value).

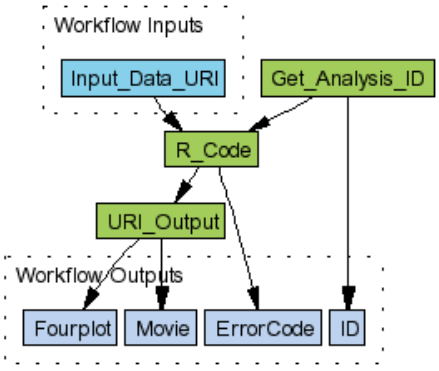

Figure 5. The Workflow in OMII Taverna

The workflow does not expose all the key interactions that take place when it is run, and so Figure 6 shows these in the form of a collaboration diagram.

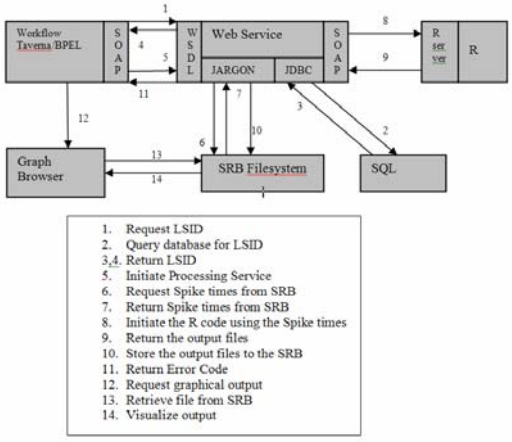

Figure 6. Workflow Collaboration Diagram

The workflow takes the URI of the spike train data as input, and calls a service – *Get_Analysis_ID* – to obtain a new URI for the analysis results. This ID is used throughout the workflow process and is returned to the user so they can view, or further process the results.

The next service – *R_Code* – receives the URI of the spike train and invokes the SJEMEA package by using the JRclient Java library to communicate with the Rserve component. This accesses the spike train data from the SRB and calls an R server (a service that can execute R scripts) to analyse the data. The results are stored back in the SRB. In the next stage, the data is accessed from the SRB and the graphs and movie produced. The images are png files and the movie is an animated gif file created using the ImageMagick [16] software which builds it up from the sequential frames of the neural activity. The graph images and movie are saved back in the SRB and from there they can be viewed after the workflow execution has been completed.

The URIs of the plots, movie and analysis results are returned as the result of the workflow, along with an error code.

5. Conclusions

CARMEN is developing a new infrastructure to support communities of neuroscientists. By building on the output, experiences and expertise of the e-science community we have been able to produce an early prototype that automates a scientific process that was previously carried out with significant amounts of manual intervention. However, accelerating research is only one benefit; another is that the primary and derived data, and the analysis services are now available for others to access and use for their own research. This provides the basis for collaboration and re-use that was not previously possible.

Novel aspects of the infrastructure include a repository for services as well as data and rich tools for searching and visualizing time-series data. Over time the prototype will be developed to incorporate other features including: fine-grained control by scientists over access to both their data and services; and, the development of metadata for neuroscience, building on the lessons learnt in bioinformatics.

Whilst the project's target is neuroscience, the infrastructure is sufficiently generic that it can also be utilized in other domains.

6. Acknowledgements

CARMEN is a large, multi-site collaboration between neuroscientists and computer scientists. The work described in this paper has benefited from discussions involving all members of the extended team but we would particularly like to thank Colin Ingram, Leslie Smith and Alastair Knowles for their work on setting the agenda the project. The scenario is built on the work of Evelyne Sernagor, Stephen Eglen and Christopher Adams. We would like to acknowledge the EPSRC for funding the project.

7. References

1. myGrid Project: http://www.mygrid.org.uk/.
2. Jackson, T., Austin, J., Jessop, M., Linag, B., Pasley, A., Ong, M., Allan, G., Kadirkamanathan, V., Thompson, H., Fleming, P.: Distributed Health Monitoring for Aero-Engines on the Grid: DAME. IEEE Aerospace. IEEE, Montana (2005)
3. BROADEN Project: www.cs.york.ac.uk/dame/broadenposter.pdf.
4. Gold Project: http://www.goldproject.ac.uk/.
5. Storage Resource Broker: http://www.sdsc.edu/srb/.
6. i3c: I3C Life Sciences Identifiers (LSIDs). http://www.i3c.org/wgr/ta/resources/lsid/docs/index.htm
7. Grimoires Project: Grimoires Registry http://www.omii.ac.uk/mp/mp_grimoires.jsp.
8. Dynasoar Project: Dynasoar http://www.neresc.ac.uk/projects/dynasoar/.
9. myGrid Project: Taverna Workflow http://taverna.sourceforge.net.
10. Signal Data Explorer: Signal Data Explorer: http://www.cybula.com/flyers/SignalData.pdf;.
11. Moses: XACML: eXtensible Access Control Markup Language, OASIS Standard Version 2.0. (2005)
12. NEReSC: Open Grid Services Architecture - Distributed Query Processing (OGSA-DQP). http:/www.neresc.ac.uk/projects/OGSA-DQP
13. Furukawa, T., Morrow, E.M., Cepko, C.L.: Crx, a novel otxlike homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. Cell **91** (1997) 531-541
14. Multichannel Systems: http://www.multichannelsystems.com.
15. R Project: http://www.r-project.org.
16. Image Magick: http://www.imageMagick.com.