The CARMEN Virtual Laboratory: Web-Based Paradigms for Collaboration in Neurophysiology

Frank Gibson[1*], Jim Austin[2], Colin Ingram[3], Martyn Fletcher[2], Tom Jackson[2], Mark Jessop[2], Alastair Knowles[1], Bojian Liang[2], Phillip Lord[1], Georgios Pitsilis[1], Panayiotis Periorellis[1], Jennifer Simonotto[1], Paul Watson[1], Leslie Smith[4].

1 School of Computer Science, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.
2 Department of Computer Science, University of York, Heslington, York, YO10 5DD, UK
3 School of Neurology, Neurobiology and Psychiatry, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.
4 Department of Computing Science and Mathematics, University of Stirling, Stirling FK9 4LA, UK.
* Corresponding author. E-mail address: Frank.Gibson@ncl.ac.uk

Abstract

Dissemination, reuse and sharing of digital resources is technically and culturally challenging in neuroscience – particularly in the area of Multi-Electrode Array (MEA) recording. Large, complex datasets are typical, with a heterogeneous range of data and code formats. The CARMEN project (http://www.carmen.org.uk/) aims to enable broad sharing or resources, through provision of a secure, online environment for data analysis, and curation of data, analysis code and experimental protocols.

1 Introduction

Multi-electrode array (MEA) based neuroscience research has matured to the point where integrated hardware and software systems for recording and stimulation can be bought off the shelf. Alongside commercial acquisition software, many neuroscientists develop custom analysis code.

Dissemination, reuse and sharing of digital resources are both technically and socially challenging. Large, complex datasets are common place with, heterogeneous range of data and code formats. Significantly, deeply ingrained social precedents impede resource sharing, despite the obvious potential for advancement both in neuroscience, and for society as a whole [1]. The CARMEN project aims to addresses these challenges, through provision of a secure environment for resource sharing, online data analysis, and precise description and curation of experimental protocols.

The benefits of sharing life-science data resources to maximise knowledge discovery are well known and documented [2]. Sharing resources can leverage community expertise, allowing others to perform different analyses on the same dataset or to perform the same analyses on different datasets, e.g. as in comparative studies across the same target animal or CNS structure.

There are several barriers to the efficient and effective sharing of resources: (a) datasets are large and complex, with no widely accepted community standards for data formats such as electrophysiology recordings; (b) experiments are not always reported consistently, or in sufficient detail to enable to third party interpretation and evaluation. Re-use of resources – datasets, analysis code or experimental protocols – is predicated on precise knowledge of exactly what these resources represent, i.e. knowledge of the study subject(s), equipment and protocols describing how the data was generated and/or processed. This is commonly referred to as metadata; data about data.

In the absence of accurate and detailed metadata resources are effectively meaningless. Annotation, the process of ascribing metadata to resources to enable future reuse, requires effort. Further, scientists may as a consequence of annotation be obliged to publish details of their protocols. Some may not wish to do this.

Ownership and accreditation are highly important. Resources, currently withheld by small groups and individuals, are globally scarce, and therefore of high perceived value. Some neuroscientists worry that they may not be appropriately credited for new results obtained using their datasets, while others are concerned that errors in their analyses may be uncovered, while others feel that resources represent competitive advantage and simply should not be dispensed to others [3].

It is recognised that the neurophysiology community is large, disparate, and global. Groups vary in size and capability, from independent researchers to highly parallel data production lines in pharmaceutical companies and research institutes. CARMEN aims to provide the basis for these diverse groups to share and repurpose resources across organisational boundaries, potentially within a single, seamless virtual marketplace. A web based platform is being developed to meet this brief.

In this paper we describe the CARMEN system and technical progress to date. We outline future work and discuss potential impact and applications.

2 Technical Progress

Initial requirements for CARMEN have been collected through extensive, iterative discussion with the project members. In summary, these requirements are: Allow "experimenter" users to describe, store and analyse data (time and image series) from various electrophysiology acquisition systems – in various proprietary and bespoke formats. Allow "analyst" users to describe, store and browse code (source and executable) for data analysis. Allow execution of code in Matlab, R, Java, Python and C/C++ languages, including parallel processes. Allow "simulation" users to describe, store and analyse data generated by simulation tools, including the NEURON and GENESIS simulators. Allow data derived from the results of analyses to be stored and bound to source data for further analysis. Allow users to specify and apply access control rights to their resources. Allow both web and client based software tools to connect to the environment securely for data analysis and visualisation. Enable data translation between different proprietary and bespoke data formats, retaining source data at all stages in translation. Support a user community that is distributed and growing, with varying data preservation capabilities and requirements.

These requirements drive the design and development of the CARMEN system.

2.1 The CARMEN Architecture

CARMEN is a Three Tier Web architecture [4], consisting of server deployments at different scales, e.g. from desktop to data warehouse. The data tier is shared between databases and a distributed (i.e. virtual) file system called the Storage Request Broker (SRB) [5]. Databases are used to store user accounts, metadata, system states and links between resources and users. Physical files, (e.g. data recordings and analysis code), are stored in the SRB. CARMEN server nodes support storage and invocation of both data and analysis services. Analysis operations may take place adjacent to data, minimising the need to transport high volume datasets. The ability to describe, search and navigate resources is provided by common frameworks for representation of data and metadata. A security framework provides the basis for groups to control access to resources, so as to enable secure collaborations, and productive competition between these collaborations. Single point access to the system is delivered via an interactive Web portal.

2.2 Representation of Metadata

Poor or insufficient metadata impedes resource interpretation, evaluation and reuse. The CARMEN consortium has defined what information and level of detail must be ascribed to a dataset that is submitted to CARMEN. This reporting consensus is called the Minimum Information about a Neuroscience Investigation (MINI) [6]. The document is separated into the following sections: contact and context, study subject, task, stimulus, behavioural event, recording, time series data. This reporting consensus is structured within the FuGE data model [7]. FuGE models common aspects of life-science experiments such as protocols, equipment and materials and is being implemented by domains such as genomics, proteomics and metabolomics. By using FuGE, CARMEN datasets can be combined with this type of information, providing a framework for integrative or system level investigation, from biological processes to neural function.

The terminology used for metadata descriptions must be associated with meaning or semantics to allow consistent description and interpretation not only for neurophysiology experiments but also analysis code and data produced. To achieve this, CARMEN aims to develop and contribute to the Ontology for Biomedical Investigations (OBI) [8]. OBI is an ontology providing consistent terminology and descriptions for life-science investigations. In this way, it is envisaged that terminology used to describe CARMEN data will be consistent and interpretable across the life sciences.

2.3 Representation of Data

A standard representation for metadata supports an environment where different experiments can be characterised in a way that is amenable to both users and computers. Raw electrophysiology recordings use a multitude of data formats – both vendor-specific and bespoke. Raw data is often unreadable without the original software or detailed knowledge of the format. This situation presents a serious barrier to collaborative neuroscience. Valuable research time is wasted decoding formats and writing data converters. To attempt to overcome these restrictions, CARMEN is developing an intermediate data format for neurophysiology recordings. The Neurophysiology Data Translation Format (NDTF) [9], provides a standard for sharing data, specifically for inter-application and data communication between analysis software developed within the CARMEN project. It provides improvements over and above existing translation formats such as NeuroShare [10], such as the ability to encode static files. Further, it is optimised for dynamic access operations such as streaming and search, allowing client tools developed in conjunction with CARMEN to support high performance visualisation and pattern searching. NDTF is a wrapped data set consisting of a configuration file and one or more ('host') data files. The configuration file is an XML [11] document which contains metadata and manages associated data. The use of NDTF provides a common mechanism for programs to interpret the semantics of data. It allows systems to 'know' which programs can be used on

which datasets, and provides a format for encoding output data.

2.4 Running Analysis Code

MEA datasets can run into terabytes presenting storage challenges and pushing desktop processing and analysis power beyond its current limits. CARMEN will help to overcome these issues by enacting analysis on the system that holds the data. Data transfers can thereby be confined to closely connected cluster machines on server nodes, enabling faster computation and shorter analysis times.

To benefit from the processing capability in CARMEN, analysis code will be uploaded and wrapped to form Web Services [12]. Web Services provide a common means of communicating with different languages and data types. Curating analysis code as Web Services facilitates the re-running and re-use of code. Early stage support has been developed for services based on MatLab, Java and R, with C/C++ and FORTRAN to be addressed in the near future.

It will be possible to enact services residing in CARMEN automatically, in user defined sequences. For example, typical MEA analyses may involve thresholding, then spike detection, and then graphical presentation of the results. Within CARMEN, the source data may be connected to the threshold service, the output of the threshold service may be connected to the spike detection service. The spike times produced form this service may form the input to the graphical presentation service, which displays the results to the user. This process of connecting services and invoking sequences by way of single commands is termed 'workflow'. A workflow is a programmatic depiction of a sequence of operations and provides a mechanism to visualize and enact processes, such as complex data analyses. Building workflows encourages users to provide modular components, rather than monolithic programs that offer less flexible opportunities for re-use. Analyses may be re-enacted by other users, by launching appropriate workflows. Moreover, the same workflow may be run on different datasets. In addition to data, code and services, workflows are valuable resources that can be shared between researchers. It will be possible to construct workflows consisting of data and services within CARMEN using a workflow editing environment, such as Taverna [13].

2.5 Security

Security and usability can be contradictory concerns. CARMEN aims be usable, intuitive and user friendly, promoting collaboration and the development of research ideas. Security (as a non functional property and also in terms of implementation) can influ-

ence the usability of a system as recent examples have demonstrated [14]. We have avoided developing a framework that compromises system usability to provide security.

Authentication and authorization are critical issues in CARMEN. Access to system resources is carried out by authenticated users in possession of appropriate authorization tokens. We are building on top of established research work [15] in the areas of Web Services to accommodate our access control and authorization requirements. Our custom security mechanisms are deployed and act as mediators between the users and system resources. We can control access and usage of resources held in CARMEN, resources deployed as web services, as well as safeguard the process of uploading and downloading data. We are exploring additional functionality such as the ability to create audit records by keeping track of access to certain resources, delegating access rights and revoking rights granted to users. Typical users will interact with the security system via the Web-based portal.

2.6 Portal

CARMEN can be accessed via a Web Portal allowing users to upload, download, view, annotate and share resources, based on the security permissions ascribed to each resource. The CARMEN portal is written in Java using the Google Web Toolkit [12], which converts the Java code to Ajax. This allows rapid prototyping, generating modern interfaces which give CARMEN the look and feel of a desktop environment. Ajax can perform asynchronous communication between browser and the Application Tier, without the need to refresh the user interface. This allows large datasets to be viewed rapidly without impacting heavily on browser performance or user experience.

3 Future Work

Provenance: Provenance provides an audit trail of analysis; what services were run, when, and what results were produced. This is similar to transcribing the process in a lab-book, except that it happens automatically. Provenance tracing mechanisms allow the derivation of resources to be captured, both for subsequent scrutiny, and to ensure appropriate accreditation and acknowledgement of originators. **Ownership:** At both social and technical application levels, the CARMEN consortium are exploring new mechanisms for publication and accreditation that truncate the resource dissemination cycle, and addressing the issues of expression and control of ownership. One proposal is the assignment of licenses to data, such as the Protocol for Implementing Open Access Data [16], or Creative Commons Zero (CC0) waiver [17]. **Networking:** A Social Networking platform constructed around data and resources may

facilitate researchers' discovery of new resources and collaborations in a manner that evokes current working practices, which would not have been possible though the traditional laboratory based working environment. **Availability:** The project aims to release a stable system to consortium members in October 2008.

4 Discussion and Conclusion

The Central Nervous System (CNS) is arguably the most complex and extensively studied biological organ. As with other complex systems, the ability to share observational resources such as data and analysis code, to chart complex dependencies, is imperative to our ability to understand its behaviour. CARMEN is disseminating descriptive frameworks for neurophysiology data and metadata, allowing resources in many different forms to be semantically linked and compared. A software system is being developed as a pilot implementation of these frameworks, allowing distributed neuroscience groups to share resources within a virtual marketplace. In future, it is hoped that CARMEN like systems may be integrated with other neuroscience databases. Globally, resource sharing initiatives like CARMEN are gaining momentum, with the FIND project (Germany), JNode activities (Japan) and emerging NSF data sharing programmes (US) spearheading advances that are highly complementary. Building on the work of these national initiatives, there is an opportunity to address global, communal metadata standards, allowing resources embedded in many different systems to be utilised in an integrative manner. The potential benefits to MEA research and the neuroscience domain are transformational; a semantic, worldwide web of data, analysis services, and collaboration and exploitation opportunities navigable from a desktop internet browser.

References

[1]   G.A. Ascoli, "Mobilizing the base of neuroscience data: the case of neuronal morphologies," *Nat Rev Neurosci*, vol. 7, Apr. 2006, pp. 318-324.

[2]   M.W. Foster and R.R. Sharp, "Share and share alike: deciding how to distribute the scientific and social benefits of genomic data," *Nat Rev Genet*, vol. 8, 2007, pp. 633-639.

[3]   J. Teeters et al., "Data Sharing for Computational Neuroscience," *Neuroinformatics*, Feb. 2008.

[4]   W.W. Eckerson, "Three Tier Client/Server Architecture: Achieving Scalability, Performance, and Efficiency in Client Server Applications.," *Open Information Systems*, vol. 10, Jan. 1995.

[5]   "Storage Request Broker"; http://www.sdsc.edu/srb/index.php/Main_Page.

[6]   Frank Gibson et al., "Minimum Information about a Neuroscience Investigation (MINI) Electrophysiology," Mar. 2008; http://hdl.handle.net/10101/npre.2008.1720.1.

[7]   A.R. Jones et al., "The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics," *Nat Biotech*, vol. 25, Oct. 2007, pp. 1127-1133.

[8]   "Ontology of Biomedical Investigations"; http://purl.obofoundry.org/obo/obi.

[9]   "The Neurophysiology Data Translation Format (NDTF) "; http://purl.org/net/carmen/ndtf.

[10]   "Neuroshare"; http://neuroshare.sourceforge.net/.

[11]   "Extensible Markup Language (XML)"; http://www.w3.org/XML/.

[12]   "Web Services"; http://www.w3.org/2002/ws/.

[13]   T. Oinn et al., "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics (Oxford, England)*, vol. 20, Nov. 2004, pp. 3045-54.

[14]   J. Wu and P. Periorellis, "Authorization-Authentication Using XACML and SAML," *Newcastle University Technical Reports*, vol. CS-TR No 907, Feb. 2005; http://www.cs.ncl.ac.uk/research/pubs/trs/papers/907.pdf.

[15]   P. Periorellis, "GOLD Infrastructure for Virtual Organisations," *e-Science All Hands Meeting*, vol. Proceedings 2006; http://www.cs.ncl.ac.uk/research/pubs/inproceedings/papers/970.pdf.

[16]   "Protocol for Implementing Open Access Data"; http://sciencecommons.org/projects/publishing/open-access-data-protocol/.

[17]   "Creative Commons Zero (CC0) waiver."; http://labs.creativecommons.org/license/zero/.