DIGITAL CURATION CENTRE

eScience Liaison

eScience Longitudinal Study No. 1

CARMEN

Code, Analysis, Repository & Modelling for e-Neuroscience

Graham Pryor, September 2008

Contents

PART 1 – The DCC and the CARMEN project

PART 2 – Designing a *lingua franca*: ontological and metadata requirements for data integration and sharing

PART 3 – Engagement with data curation issues, drivers and constraints

PART 4 – Concluding perceptions

PART 5 – Conclusions and acknowledgements

Appendices

Digital Curation Centre eScience Longitudinal Study No. 1

CARMEN [Code, Analysis, Repository & Modelling for e-Neuroscience]

Legal Notices

Catalogue Entry

Title	CARMEN [Code, Analysis, Repository & Modelling for e-Neuroscience]
Series Title	DCC eScience Longitudinal Study No. 1
Creator	Graham Pryor
Subject	Data curation, data sharing, eScience, human infrastructures, metadata, neuroscience, neurophysiology, repositories, research practice, virtual laboratories
Description	The CARMEN pilot project seeks to create a virtual laboratory for experimental neurophysiology, enabling the sharing and collaborative exploitation of data, analysis code and expertise. This study by the DCC contributes to an understanding of the data curation requirements of the eScience community, through its extended observation of the CARMEN neurophysiology community's specification and selection of solutions for the organisation, access and curation of digital research output.
Publisher	Digital Curation Centre, Edinburgh
Date	September 2008
Type	Text
Format	PDF Content v. 1.4
Resource Identifier	ISSN 1759-5452
Language	English
Rights	© Digital Curation Centre

Part 1 – The DCC and the CARMEN Project

1. Introduction

This report is the summation of an extended series of observations of the CARMEN pilot project.[1] The study of CARMEN, which stands for Code, Analysis, Repository & Modelling for e-Neuroscience, was conducted between October 2007 and August 2008[2]. It comprises the first in a series of longitudinal studies undertaken for the Digital Curation Centre (DCC) eScience Liaison programme.

1.1 Background to the longitudinal study

The overarching aim of the CARMEN project is the creation of a virtual laboratory for neurophysiology, which will enable the sharing and collaborative exploitation of data, analysis code and expertise that are not physically co-located. A significant development project, CARMEN exhibits what have come to be regarded as the key characteristics of eScience:

- **cross-domain collaboration in which the conduct of research is dependent upon the electronic sharing of data;**
- **a research environment enabled by the interconnection of computers;**
- **the use of ubiquitous networks to provide high speed data transmission across a geographically distributed community.**

A principal commitment within the DCC's Phase 2 programme is to understand the data curation requirements of the eScience community, with the dual aim of promulgating good practice or proven solutions amongst that community and informing the orientation of DCC tools and services around actual needs and requirements.

From exploratory discussions with contributors to the project proposal, CARMEN was identified as an important opportunity to observe the criteria and processes used by active researchers and their support teams. Given what was known about the complexion of the CARMEN consortium membership, the group appeared likely to typify conduct within the neurophysiology domain in the manner of its approach to the specification and selection of solutions for the organisation, access and curation of digital research output.

It will be recognised from the project proposal[3] that the value of CARMEN as an eScience test bed was neither predicated on nor limited to the provision of technical solutions for the curation and preservation of data. Instead, the driving aim of the CARMEN pilot project has from its conception been described as the achievement of a fundamental step change to the manner in which neurophysiological research is conducted. This vision was expected to give rise not only to a considerable technical challenge but also a broad range of cultural and behavioural issues.

Tracking these more fundamental aspects of the CARMEN programme has been of no less interest to the DCC, since it is accepted that the individual research context and culture exercise considerable influence over the manner in which data is generated, managed and shared.

In any case, and in support to its stated aim, CARMEN is a considerably datacentric initiative, having a major and consistent focus on the organisation and value of neurophysiological data as a function of improving the science. This was made clear in a paper delivered to the 2007 eScience All Hands Meeting, when it was claimed that data from experimental neuroscience is 'difficult and expensive to produce', yet 'it is rarely shared and collaboratively exploited'.[4] Resonating within the evidential foundations to that claim were five specific problem characteristics, described in the paper as the key challenges to which the CARMEN project would seek to provide solutions and accommodations. Four of them pointed explicitly at data issues:

[1] http://www.carmen.org.uk/

[2] Some additional material was identified in September, during the insertion of revisions to this report

[3] Original project proposal: http://www.carmen.org.uk/about/CARMEN-PROPOSAL-FINAL.pdf/view?searchterm=project

[4] Paul Watson et al, *The CARMEN Neuroscience Server*, http://www.allhands.org.uk/2007/

1. data that are difficult to integrate because they are generated in high volumes, frequently using proprietary software, and bearing heterogeneous characteristics as a consequence of locally provided descriptions and data curation;
2. a dearth of analysis techniques that can be applied across neuronal systems, which as an ambition is frustrated by an absence of accessible data that can support the development of analysis techniques;
3. the lack of any structured approach to the curation and optimisation of data across the domain;
4. An absence of sustained interaction between distinct and disparate research centres with the potential to share and exchange complementary expertise;
5. A limited understanding within the community of the informatics solutions that could be applied to the integration of data.

The resolution of these challenges was predicted to have a far reaching impact, with the potential to benefit a much larger research community than is represented by computational and experimental neuroscience. Associated disciplines that were also expected to gain advantage from the project included:

1. Computer science (neuromorphic and neuromimetic systems)
2. Nanotechnology (neuronal interfaces, neuroprostheses, biosensors)
3. Electronic engineering (design of multi-sensor systems)
4. Informatics (data handling)
5. Pharmacology (*in silico* drug development)

1.2 Methodology and scope of the study

A longitudinal study is usually described as a programme of correlational research, involving repeated observations of the same groups of items or individuals over a length of time. This DCC study of CARMEN was undertaken over a period of only eleven months. It was not afforded dedicated effort during the whole of that elapse of time, but through periodic and sustained interactions with the project it was possible to track the emergence of a range of technological and organisational approaches to the principal data handling and curation issues known to be frustrating data sharing and integration in the neurosciences.

Observations of the project's progress were made through participation in CARMEN consortium summits; through individual meetings with technical project team members and with research members of the consortium; through telephone interviews; from an analysis of documents and progress reports produced by the project; and through engagement with the project's open email correspondence and online blog.

The assessment of how significant a step change was achieved in establishing appropriate and robust mechanisms for the curation of neurophysiological data was therefore dependent not only on measuring the convergence of the project's objectives and its final deliverables, but also on recording the incidence and nature of development and progression. Subsequently, the report narrative attempts to capture key episodes in the evolution of the project. Consisting of discrete sections that address the pivotal features of the CARMEN service, such as its metadata architecture or data sharing principles, themes are reproduced to reflect the sequence in which they were developed during successive meetings and dialogues.

Whilst CARMEN is a large, multi-site collaboration between computer scientists and neuroscientists, the identification of generic benefits to the wider eScience community has also formed a crucial aspect of the study.

Data specific issues that suggested themselves as worthy of attention by the DCC study included:

- How the CARMEN eScience infrastructure approaches the coherent integration of diverse data;
- What level of facility is achieved in the provision of accessibility and user interaction;
- The nature and effectiveness of protocols for managing access to data and services;
- Underlying preservation, security and rights-management issues;

- The design of mechanisms for the creation, allocation and development of standard metadata, and how the metadata are sustained in relation to large, complex and dynamic data sets;
- The extent of provision for legal compliance, including intellectual property rights (IPR);
- The provision for migration, archiving and the handling of proprietary products/tools;
- System/service longevity: the post-project strategy for sustainability, development and return on investment;
- The success of the neuroscientist/computing scientist partnership in delivering a working solution, with observations on the range and availability of expertise required for its successful delivery.

In the eventuality, not all of these issues could be explored in depth, particularly where their resolution was reserved for later in the project. Neither is an in-depth explication given of the CARMEN technology platform, system structure, or repository design, since the focus of this study has been confined to issues concerning data organisation and management. An introduction to the service infrastructure is given in Part 3 of this report, but more detailed information about the CARMEN technology environment may be sought directly from the CARMEN project via its Web site[5] and through papers published by consortium members.

Most fundamentally, the persistent thread of interest for this study has been the attitudes, needs, processes and relationships of experimental neuroscientists in their generation and use of data, and the impact these key human factors have on effective data curation.

Funding for the current phase of the DCC was granted with the proviso that scientific data would be made a priority focus of attention. This study of CARMEN is one filament of that new focus. There were two other conditions of grant: that the DCC should develop the skills level of support staff in order to ensure strength in depth and that its potential for transition from a development project to a service should be explored. Both conditions are relevant to the CARMEN project and the sharing of experiences and speculation on skill and service development became an unplanned addition to the study.

[5] http://www.carmen.org.uk/

2. CARMEN project rationale

CARMEN is a £4.5 million, four-year eScience pilot project funded by the UK's Engineering and Physical Sciences Research Council (EPSRC). Organised as a consortium of twenty academic investigators, and representing eleven universities, it commenced on 1st October 2006. The driving aspiration shared by members of this consortium is for CARMEN to make possible an entirely new and more effective means of conducting neurophysiological research. It is planned that this will be achieved through the introduction of technology-enabled methods for sharing experimental data.

2.1. Constituents and culture

The eleven institutional partners in the CARMEN consortium are the Universities of Cambridge, Leicester, Manchester, Newcastle, Plymouth, St Andrews, Sheffield, Stirling, Warwick, York and Imperial College. Commercial partners include Cybula Limited and Lectus Therapeutics. The project's Principal Investigator (PI) and coordinator is Professor Colin Ingram, Director of the Institute of Neuroscience at the University of Newcastle.

The project investigators are senior academic staff working in neurophysiological research, coming from backgrounds in the disciplines of neuroscience, neurobiology, computing science, computational neuroscience, applied mathematics and psychology. Researchers in Germany, Japan and the USA are also associated with the project.[6]

Research into the central nervous system is believed to involve in excess of 100,000 neuroscientists worldwide. At a global level, sufferers from disorders to the central nervous system number an estimated 450 million[7], yet effective treatments remain limited. Progress in unravelling meaning from the billions of sensory inputs to the central nervous system is understandably slow, when research experiments themselves generate correspondingly massive and complex data sets.

Neurophysiological research is critically dependent upon the collection and analysis of these data, and considerable effort is expended using data captured from neuronal systems[8] to develop models that will explain not only the processes that gave the data their characteristics, but also the functions expressed by those characteristics (typically, behaviour and thought). The techniques employed in capturing and analysing these data are not only time-consuming; they are also difficult to use, expensive and extremely diverse.

Data from neuronal systems are frequently derived using multichannel electrical recording (i.e. the use of arrays of electrodes to record directly from neuronal tissue). Some researchers use ion-sensitive fluorescent dyes, highlighting chemical changes in the tissue which can be captured using optical imaging (e.g. high frame-rate photography). Between those two main categories there exist a large number of specific techniques, each having its own unique advantage (e.g. it may produce a high time resolution, or alternatively a high spatial resolution). The heterogeneity of the techniques applied is further compounded by the range of models built to test and analyse the data recorded. As described in the CARMEN project overview, these cover a wide spectrum, running from 'the detailed modelling of membrane-embedded ion channels and neurotransmitters to compartmental neural models, through models of small neural networks, to larger models of many thousands of neurons'.[9]

Historically, data are shared amongst members of a laboratory. Laboratory groupings typically include preferred computational analysts and/or modellers, who apply mathematics to the

[6] A full list of consortium members may be seen at http://www.carmen.org.uk/people

[7] Figures supplied courtesy of the CARMEN Project factsheet at http://www.carmen.org.uk/about/carmen-factsheet.pdf

[8] Simply described, neuronal systems are networks of brain cells; in the research environment, these networks can be both live and cultured. Experimental neoroscience is undertaken *in silico*, *in vitro* and with living organisms (including humans)

[9] http://www.carmen.org.uk/about

understanding of the neuronal systems that the data represent. In this sense, the science is distributed: the data producers and analysts attached to a laboratory are usually physically remote from one another, even in the very early stages of the data lifecycle.

As both data producers and analysts have very specific motivations, it takes a long time for mutual 'attraction' and trust to develop to the point at which data sharing can take place. Therefore, laboratory groupings typically contain no more than one or two analysts. It can be analogised that this monogamy of data producers and analysts is not optimal to scientific conception. To continue the analogy, CARMEN aims to provide an environment where greater promiscuity can be explored, allowing a spectrum of partnerships at different levels of remove to make better use of available data.

The implications for data curation are profound. Laboratory groupings perceive themselves as islands, and therefore invest minimal curatorial effort into making themselves 'attractive' to the induction of new members. The converse may also be true. Barriers to integration mitigate competition and safeguard individuality, and may therefore be actively upheld. Data sets are organised idiosyncratically, and typically, a translation step is required every time a migration route is established between one laboratory grouping and another. Further, data are seldom publically archived in a format that is widely accessible, as it is assumed that everything that passes along the migration route will be maintained in duplicate at both ends.

While the rationale for challenging these precedents is obvious, CARMEN must engage its community in managed transition, rather than imposing a rate of change that is intolerable. This offers the most expedient path to improved curation.

2.2. Aims and objectives

Using the GRID,[10] the CARMEN team aim to replace these historical processes with technology-enabled services, providing experimenters with a common structure for archiving their data sets and making them widely accessible for exploitation both by authorised modellers and those researchers engaged in the development of new explorative algorithms.

Perhaps most significantly, within the project it has been recognised that experimental data sets are virtually useless when unaccompanied by accurate descriptions of the experimental conditions pertaining. Consequently, the development of an appropriate set of metadata/representation information has, from the outset, featured as a critical element in the project work plan. Without this planned augmentation of the experimental data due for upload to CARMEN, the primary goal of facilitating greater opportunities to collaborate more widely and persistently could not be achieved, since the effective sharing of data depends upon rendering it in a referenced and accessible form, using terminology that will be comprehensible to all those working within the domain's knowledge base.

Additionally, the project seeks to provide a range of integrated and co-ordinated services for neuroscience data, enabling neuronal signal detection, sorting and analysis, as well as visualisation and modelling. Originally, these included plans to support experimental optimisation by streaming data directly into CARMEN for real-time analysis. This has been put on hold, as the technical challenges are currently beyond the scope of the project.

These aims are consistent with the project's goal of catalysing a step change in research practice in this area of neuroscience, the predicted outcome from which will be the derivation of best return on the significant investment that is being made in research into the human brain.

[10] The UK National Grid Service is described at http://www.grid-support.ac.uk/

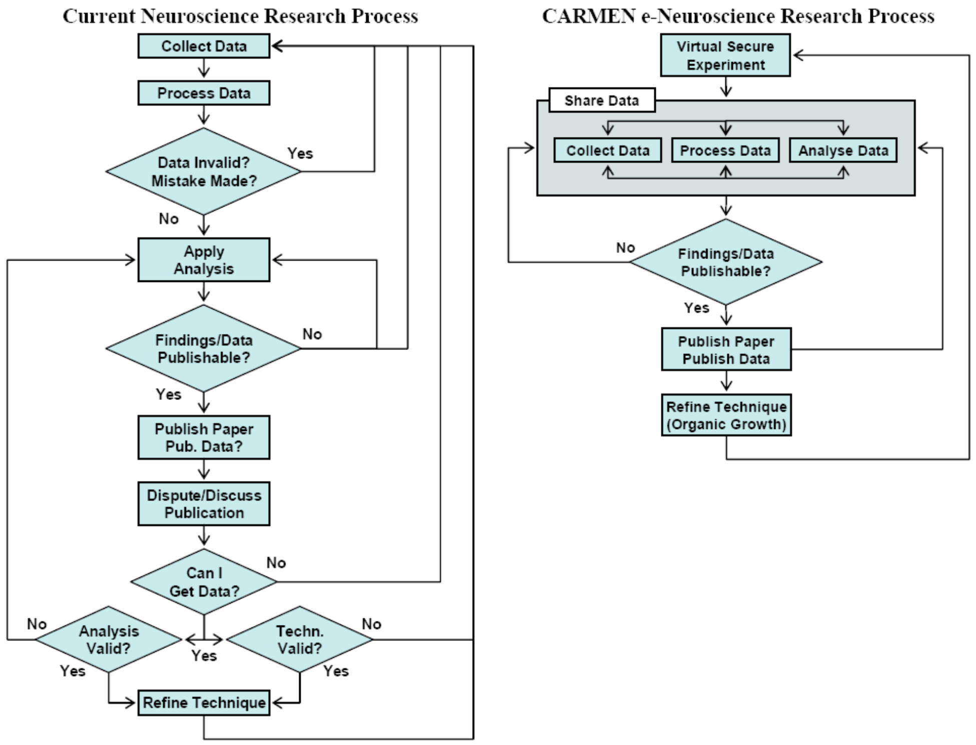

Current Neuroscience Research Process
Collect Data
Process Data
Data Invalid? Mistake Made?
Yes
No
Apply Analysis
Findings/Data Publishable?
No
Yes
Publish Paper Pub. Data?
Dispute/Discuss Publication
Can I Get Data?
No
Yes
Analysis Valid?
No
Yes
Techn. Valid?
No
Yes
Refine Technique
CARMEN e-Neuroscience Research Process
Virtual Secure Experiment
Share Data
Collect Data
Process Data
Analyse Data
No
Findings/Data Publishable?
Yes
Publish Paper Publish Data
Refine Technique (Organic Growth)

Figure 1 – Traditional Neuroscience Research Process and CARMEN-enabled Research Process[11]

In the context of this DCC study of CARMEN, it is to be noted that whilst the project proposal may declare metadata to be 'critical to ensure that the stored data can be discovered and understood',[12] the project's principal aims and objectives are not described as the implementation of techniques and standards for data curation. CARMEN is an eScience project with a firm focus upon the scientific understanding of neuronal activity, encompassing ionic conductances, synaptic and spiking activity and network dynamics, and with a programme to engineer, utilise and refine a scalable eScience architecture to serve this research.[13]

Nonetheless, a concern to manage data effectively is more than implicit to the longer term benefits to be derived from the project. In the project proposal, whilst the CARMEN consortium is depicted as pioneering a rationalisation of research practice by seeking to embed organised data and software service resource sharing, critical benefits are forecast to include (a) close collaboration between data collectors and analysts during experimentation; (b) cross-fertilisation and convergence of expertise between areas of specialisation; (c) ***diligent and long-term curation of data and tools***, and; (d) ***optimal re-use and integration of data***.[14]

For the DCC, therefore, a genuine and appropriate interest in CARMEN lies in how well it meets the eScience challenge defined in the project proposal:

'to engineer an extensible Grid-based system that is: (a) generic to data-intensive neuroscience; and (b) supportive of very dynamic, flexible, integrated and secure deployment of domain semantics at the metadata, data and service levels'.[15]

[11] http://www.carmen.org.uk/about/CARMEN-PROPOSAL-FINAL.pdf/view, 1.2, Research Benefits and Deliverables
[12] CARMEN-PROPOSAL-FINAL, 2.2, Metadata
[13] CARMEN-PROPOSAL-FINAL, 1.2, Research Benefits and Deliverables
[14] CARMEN-PROPOSAL-FINAL, 1.2, Research Benefits and Deliverables [bold italics shown in this report only]
[15] CARMEN-PROPOSAL-FINAL, 1.1, International Context and Drivers

3. CARMEN project management, methodology and dynamics

At a strategic level, a Steering Committee comprising a representative selection of consortium members convenes on an annual basis, to advise and externally appraise the project.

At an operational level, CARMEN is managed by an Executive Committee, which is coordinated by the Project Manager, Alastair Knowles. It includes the Principal Investigator and senior computing scientists and bioinformaticians. The Project Manager was not appointed and in post until August 2007, after ten months of the project had elapsed, during which time CARMEN business had been progressed by members from the Executive Committee.

The project has been organised into seven teams to deliver the following work packages:

WP0 Virtual laboratory infrastructure; user requirements gathering and support
WP1 Spike detection and sorting
WP2 Information theory methods for spike trains, continuous waveforms and image series
WP3 Search algorithms for ion channel conductance parameter determination in cellular membrane models
WP4 High performance template and context based signal data search
WP5 Synchronous neural activation in motor disorders; causality analysis and visualisation; real-time data streaming
WP6 Neural activity dynamics in the neonate retina; models of activity spreading in epilepsy; Bayesian correlation methods

A Work Package Coordinators Committee managed by Alastair Knowles meets on a monthly basis to review progress across all of the work packages. The minutes of these meetings are shared with the consortium through publication on the CARMEN Web pages.

WP0 has provided most of the focus for the DCC study, together with the activities of the Metadata Working Group, which is led by Frank Gibson, Research Associate (Computing Science) at the University of Newcastle. WP0 will deliver the hardware and software to enable the neuroinformatics services described in WP1 to WP6.[16]

Since the arrival of the Project Manager, the importance of good communication to the coherence of a distributed project has been recognised and addressed. A new version of the CARMEN Web site went live midway through the DCC study, having been designed to allow consortium members to contribute news items, events and blog posts, and to upload their CARMEN related publications. With the aim of ensuring universal accessibility and inclusivity within the consortium, it includes an online manual for use of the site. Twice-yearly consortium meetings are held at which project progress is reported, at a work package level. These are well attended events at which matters arising during the previous six months are discussed and resolved, or actions nominated, and they serve as an opportunity to demonstrate aspects of the system's emerging functionality, to air new discoveries and to test potential solutions in open forum.

The nature of a geographically distributed development project will always introduce particular dynamics requiring attentive and articulate management. For CARMEN, that distribution has crossed the cultural boundaries of a number of individual academic institutions, as well as having produced a project team from a range of distinct academic disciplines. Further, the creation of a fully functional software product, albeit one that is designated a pilot, is not normal business for academics in many of those disciplines. Even for the technical developers from computing

[16] Work packages 1 to 6 are described by L S Smith et al in *The CARMEN e-Science pilot project: Neuroinformatics work packages*, http://www.allhands.org.uk/2007/

science, working on a large multi-site development would be expected to present a number of hitherto unfamiliar challenges.

A major dynamic of the project concerns the tension between theoreticians and practitioners, which affects how the project addresses access, the management of data and the provision of tools. The domain is traditionally very inward-looking, supporting a culture of communication on a one-to-one rather than a one-to-many basis, which process requires a granting of trust before data will be shared. Relationships are defined by progress through a closed network, from which its members obtain confidence. Contrary to rational expectations, the CARMEN consortium agreement does not bind its members into any consensus concerning agreed levels of openness; indeed, in the prevailing culture, attempts at a more authoritarian or overtly directed approach could prove counter-productive.

It is, therefore, not inappropriate to regard the CARMEN Project as ambitious in terms of both its organisational and technical context. Both aspects are evident from the official description of the virtual laboratory that is currently under construction.[17] As a federation of server nodes, CARMEN will allow distributed data to be stored locally to their point of acquisition, with data sets constructed across multiple nodes. Analysis codes for the interrogation or modelling of data sets will be uploaded and executed on the individual nodes, so that derived data sets need not be transported over low bandwidth connections. Data and analysis codes are to be described by structured metadata, which will provide an index to enable searching, annotation and audit over workflows leading to scientific outcomes. Users of the CARMEN service will access these distributed resources through a web portal emulating a PC desktop.

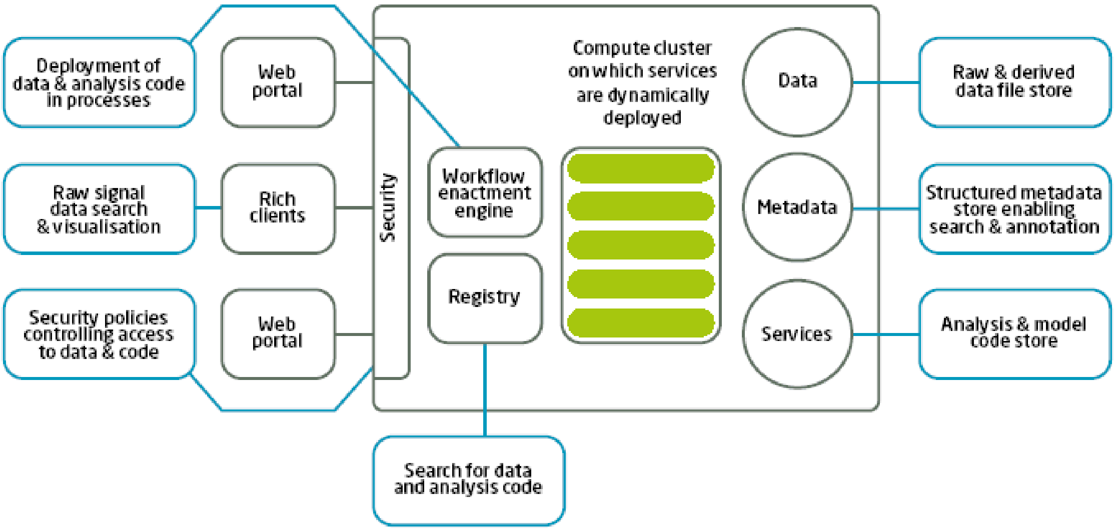

Deployment of data & analysis code in processes
Web portal
Raw signal data search & visualisation
Rich clients
Security policies controlling access to data & code
Web portal
Security
Compute cluster on which services are dynamically deployed
Workflow enactment engine
Registry
Data
Metadata
Services
Raw & derived data file store
Structured metadata store enabling search & annotation
Analysis & model code store
Search for data and analysis code

Figure 2 – CARMEN system schematic, ***courtesy of the CARMEN Project***

CARMEN may also be considered ambitious for introducing new technology solutions across a considerably diverse constituency of researchers, where will be found an array of technology literacy as well as some methodological idiosyncrasies. To achieve an effective CARMEN community, essential training and guidance has not been overlooked, with provision taking the form of reference documentation, one-to-one handholding and formal training events.[18] Active support will be provided to system users via a telephone contact. In terms of system design, the approach taken is one that aims to reproduce familiar practices and tools, such as the use of

[17] http://www.carmen.org.uk/about/carmen-factsheet.pdf
[18] The programme for a users workshop on Web services and workflows is provided as an example at Appendix 2

simple drop-down lists for the selection of metadata, thereby reducing the number of situations requiring assistance.

The critical assumption that underlies the CARMEN product is that neuronal data from members' research will be made more available. CARMEN is not merely a repository in which individual researchers or research teams can safely and covertly store their data, but an effective means of sharing and re-using those data. This is the essence of CARMEN. Yet, whilst one may have expected it to have been fully accepted by consortium members who, by joining, had committed to the vision and goals of the Project, it has proved the aspect most likely to produce resistance. As a consequence, and in several areas of the systems development, members of the WP0 technical development team have found themselves engaged in the accommodation of individual concerns raised by this fundamental cultural change. Optimistically, one might predict the outcome as a product designed in all respects to be responsive to user requirements. The potential risks from endless compromise are, however, equally predictable.

4. CARMEN data service infrastructure

The tangible deliverables from CARMEN will comprise data and metadata repositories for experimental neuroscience data and a suite of neuroinformatics services that will supply and can be operated on the data provided by experimental neuroscientists and neuroinformaticians.

Initially, there will be two repository sites, one at the University of Newcastle and the other at York. Although they will have the same infrastructure, data will not be mirrored across the two sites. This environment of federated data will be accessed by consortium members via a single interface (i.e. through a process of dynamic deployment, users will be routed to data irrespective of its location).

Together, these repositories have been named the CARMEN CAIRN (CARMEN Active Information Repository Node), which is to be considered as the original nexus of a distributed whole. In the future, assuming a larger proportion of the global neuroscience community eventually subscribes to CARMEN, it is envisaged that geographically dispersed groups would create their own CAIRNs, and that the CARMEN integration function would be scaled up to support the presentation of data and services across this expanded family of CAIRNs, as if they were a single entity.

Data uploaded to the CAIRN by experimental neuroscientists will include both raw voltage signal data that has been recorded in the laboratory and images captured from neuron and neuronal network activity. From prior experience it is known that these will represent a very high volume of data, with an initial storage requirement predicted at 50 terabytes.

Users of CARMEN will interact with the CAIRN via a Web portal, which will serve as a conduit for uploading their data, for annotating it with metadata, and as a facility for locating and browsing data within the CAIRN. The portal will also provide the means to create, run and monitor workflows. These are the visible deliverables with most immediate and conspicuous impact for the participating neuroscience community, but a more fundamental consideration of the data has been intrinsic to the CARMEN project since its inception. In their paper to the 2007 All Hands meeting, Paul Watson *et al* referred to the CAIRN as a repository for the long-term storage and curation of both data and analysis services, which services would typically support spike identification, statistical analysis and visualization:

"Scientists upload their services in a deployable form into the CAIRN where they are 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 for processing (something that will be very, perhaps prohibitively, expensive when large data sets are involved)."[19]

Interestingly, in a paper to the Google Conference on Scalability[20], the CAIRN architecture is portrayed as a 'science cloud', where it is ascribed a considerably generic complexion in terms of both technology and services. Domain specific neuroinformatics services and content are built on top of this assembly of generic solutions, a level above the middle layer of core e-Science cloud services such as workflow and metadata management, opening the way in the long-term for the option of utilising external pay-as-you-go cloud services to provide basic storage and processing facilities at the bottom layer. The potential value of CARMEN as a generic service model is therefore evident.[21]

In terms of providing for long term preservation of the data itself, it had been planned that the primary data would be held unchanged and unprocessed in a Storage Resource Broker (SRB) file

[19] Paul Watson et al, *The CARMEN Neuroscience Server*, http://www.allhands.org.uk/2007/

[20] Paul Watson & Jim Austin, *CARMEN: a Scalable Science Cloud*, Seattle, 14th June 2008

[21] A representation of the CARMEN eScience Cloud is provided in Appendix 3

store[22], whilst derived data would be stored in a database that could be operated on using CARMEN's full range of service functionality. This separation was seen as underscoring the WP0 project team's concern to retain the integrity of the source data. Over time, whilst the metadata describing this data would be developed, the source data itself would be effectively 'frozen' in an attempt to preserve the science. Subsequently, it has been decided to keep both raw and derived data in the SRB, with only the metadata held in databases, since the option of using an RDBMS for derived data proved too complex.

These commitments received support at the October 2007 consortium meeting when, at the conclusion of a critical analysis of the need to preserve metadata, it was agreed in plenary never to remove existing terms but only to add new ones. This decision was perceived as a means of accommodating new levels of understanding as they continue to emerge in the field of brain research, progress in which is marked by the creation of new subdivisions, and hence more detailed descriptions of areas of the brain that were previously described at a higher and more general level. The consequence here is a need both to add new and more detailed terms as well as to expand, augment and clarify existing, more high-level terminology.

Associated with that assessment is the knowledge that the tools used do not remain static. There is a recurrent motivation to return to earlier data and, using new tools, produce new results and new publications. Current measurement techniques are coarse, relying on variable and unsophisticated technology; the brain is still largely an unknown, and extracting data from an organic source can be an imprecise science. However, scientific methods and the technologies used will advance, with the introduction of new techniques that could enrich the interpretation of data captured previously – so long as it has been appropriately preserved and curated.

The October meeting was already clear which data it would be essential to retain,[23] and members could also support the metadata strategy with examples of data platforms and tools having changed (e.g. reference was made to a current research requirement to access and use data that was stored twenty years ago, where reinterpretations are producing new metadata).

From an entirely different perspective, other questions relating to arrangements for data retention and preservation brought into the open a proposal by the WP0 team that where data is not used or is little used it would, after two to three years, be moved to an archive (most probably a tape store). The reasons for this are more to do with the cost of providing and optimising expensive disk space than issues of data integrity, a consideration as pertinent to an established repository service as to this pilot project.

[22] Produced by the San Diego Supercomputer Center, the SRB software infrastructure supports shared collections that can be distributed across multiple organisations, locations and storage systems

[23] Time series data; spatial data providing the location of neurons; the location of Multi-Electrode Arrays (MEAs); visualisations of data sequences

PART 2 – Designing a *lingua franca*: ontological and metadata requirements for data integration and sharing

5. Demands for a common language

The resources and people involved in neurophysiological research are divided not only by distance and organisational boundaries but also by a broad range of applications, models and schema. The neuroscience community embraces a number of individual disciplines and the meaning of terminologies and nomenclature is not consistent across that community. It was essential for the CARMEN team to bridge the disparate 'dialects' by creating or adopting an ontology that would sustain the shared and consistent understanding of terms used, and a metadata architecture that would support the combination of different data sets, whilst also providing a means of discovery, interrogation and access via any legitimate route to the CARMEN data resource.

Lack of support for the concept of a shared language was one hurdle the Metadata Working Group did not have to face. At a relatively early point in the project, during the October 2007 consortium meeting, members of the CARMEN community expressed their concern that the construction of a metadata schedule was not being addressed with sufficient urgency. This proved to be a false perception, as was made evident when Frank Gibson, the Technical Moderator for metadata issues, responded by explaining the set of principles that had been adopted, as well as describing a series of activities that was already underway.

However, following a lengthy discussion of the project's metadata requirements, initiated from the floor, it became clear that the real concern of the CARMEN community was the desire for a significantly greater level of involvement in the selection and definition of metadata to be used with the system. This swell of opinion seemed to be not so much an exhibition of scepticism concerning the WP0 technical team's methods or expertise, but more an expression of 1.) the level of ownership that Carmen members feel towards their experimental data and 2.) their concern to achieve an optimum fit with the actual working requirements of the domain. Whilst the latter concern was to be regarded as encouraging, the issue of data ownership would recur as a potential threat to project success.

6. Principles and standards for data description

The principles that had been adopted and the rate of progress that was described to the October 2007 consortium meeting included:

- adoption of the FuGE[24] (Functional Genomic Experiment) standard as the model for describing common components such as materials, data, protocols, equipment and software, which can be extended to develop modular data formats with consistent structure. As a framework for describing complete laboratory workflows, FuGE allows the capture of additional metadata that give formats a context within the complete workflow;
- use of an XML structure to represent the individual elements of neuroscience experiments;
- reference having been made to relevant existing or developing ontologies, in order to explore the use of terms in fields related to neuroscience (these included the OBI Ontology project, which is developing a standard for biomedical investigations[25]);
- confirmation that the CARMEN metadata architecture had been uploaded to the project's Web site for comment;
- a plan to seek agreement with the CARMEN community on what is understood by specific terms in use within the discipline;
- the provision of use cases from the ongoing requirements gathering programme that will assist in the definition of terms;
- the Technical Moderator's critical examination of existing systems in neuroscience and related fields (e.g. http://neurodatabase.org) to identify the principal issues that must be addressed when building a metadata schema.

Indeed, the provision of good metadata was always accepted by the WP0 development team as singularly critical to enabling the discovery of data stored in the CAIRN. It was also understood that to provide a fully functional service it would be essential for the metadata not only to identify the data but also to facilitate their interpretation, through descriptions of the experimental context in which data were gathered. As described to the 2007 All Hands meeting, it was planned that

> **'the tools provided for importing data into the repository will enable the user to specify and re-use 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 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.'**[26]

[24] See http://fuge.sourceforge.net/

[25] OBI aims to produce a set of 'universal' terms for use across various biological and technological domains and domain-specific terms relevant only to a given domain. This ontology will support the consistent annotation of biomedical investigations, regardless of the field of study, and will model the design of an investigation, the protocols, instrumentation and material used, the data generated and the type of analysis performed - see http://obi.sourceforge.net/

[26] Paul Watson et al, *The CARMEN Neuroscience Server*, http://www.allhands.org.uk/2007/

7. ‘Metadata of the week’

To meet his stated aim of seeking community agreement on the meaning of specific terms, the Technical Moderator launched an experimental process described as ‘metadata of the week’ on the first Friday following the October 2007 meeting. Introduced in an email to the whole consortium, he proposed to circulate a ‘term of the week’, which all consortium members were encouraged to consider and discuss, eventually agreeing a preferred description. Ostensibly, this unusually democratic process was intended as an effective means of collecting relevant terminology with its appropriate definition, and it would run in parallel to the scheduled requirements gathering programme, which had already mobilised a series of interviews with participating neuroscientists. Members were invited to comment on this as a good or bad idea; from the outset there were no dissenting voices.

The first term to be circulated was one that had been judged to be in very common use within the consortium. As an opening *Aunt Sally*, the term ‘spike sorting’ was accompanied by a simple definition: ‘a process where big sharp spikes are separated from small blunt spikes using something called Matlab’. There followed a lively email discussion of spike sorting as a classification procedure, and by the end of that afternoon it was reported that ‘this very useful exercise’ had produced five definitions for spike sorting. Consortium members were then asked by the Technical Moderator ‘are they all correct or does one stand out above the rest?’.

So far, the process had demonstrated obvious value in encouraging a sense of inclusivity, as well as the potential benefit from drawing on expert knowledge.

The initial definition that had been offered for spike sorting had set very wide boundaries, in that it did not restrict the context to neuronal data, whilst at the same time it managed also to appear over-specific, since it confined the experimental environment to MatLab[27]. Perhaps because of this intentional absence of rigour, by the following Monday a further round of discussion had produced definitions for three related terms - a spike, spike detection and spike sorting - where these had been unbundled as serial properties from within the original term. With this success, consortium members were then asked to suggest which other terms would benefit from the same collaborative process, an approach that might be judged to imply a certain lack of structure. Furthermore, since the process of delivering a metadata service was expected to be complete by mid-2008, reliance on weekly consultations looked dangerously over-ambitious, for with one term on offer each week they would produce in the region of just thirty-six definitions.

The second term to be offered was ‘electrode’. This generated an even greater volume of debate, with differences of opinion caused by variations in the discipline background of respondents. Eventually, one member with expertise in electronics proposed a definition that he had taken from a dictionary, which the technical moderator normalised and announced to be the preferred definition: ‘An electrode is an electrical conductor device through which an electric current enters or leaves a substance (or a vacuum). For example, a substance of which electrical characteristics are being measured, used, or manipulated. Electrodes can be used to detect electrical activity such as brain waves’. There followed even more heated debate over the legitimacy of including the role of an object in its definition, with the moderator claiming that it was invalid and a senior member of the WP0 team adamant that it was essential.

It seemed pertinent at this point to reappraise the rationale for such a communal approach to the definition of terms. Since the original definition of ‘electrode’ that was floated to the consortium had been almost identical to one provided in *Wikipedia*, and with the finally preferred term having been derived from a dictionary, there was a risk that the exercise could appear spurious. Certainly there was evidence that it was becoming a source of irritation, for towards the end of the electrode debate the Principal Investigator asked to be taken off the circulation list, his mailbox having been overwhelmed.

[27] A proprietary numerical computing and programming language created by *The Mathworks* and used extensively but not exclusively in computational neuroscience

Subsequent private correspondence with the Technical Moderator confirmed that the *Metadata of the week* exercise was not essential to the definition of terms. Prior experience in the proteomics domain and with OBI Ontologies (the standard for biomedical investigations) had already enabled the profiling of a metadata schema for CARMEN, and the process now underway was intended more to highlight to the consortium the necessity of agreeing a shared understanding of the terms that would be used, rather than build the metadata structure from scratch. Its value was therefore perceived from the level of engagement it engendered, where this contributed to the development of trust in the metadata that would be essential to CARMEN's effectiveness. Most of the terms themselves were expected to come from the requirements-gathering exercise which was already underway across the individual consortium sites, as well as from the subsequent development of use cases. At the same time, as metadata can never be complete, for it will always lag behind new experimental techniques or procedures, this communal analysis of terms had served a valuable purpose in revealing the extent to which metadata is to be considered dynamic, with the corollary that the producers of research data should expect to engage with it on something approaching an organic basis.

In terms of a methodology for building a schema, the Technical Moderator was categorical that the focus for the metadata of the week exercise was principally advocacy for the natural language definition of terms. There was never an intention to build a schema over email. The established schemas upon which the Carmen terms would be developed had already been identified and within less than six months a formal approach to handling information about neuroscience was to be published. Later, in interviews[28] with some of the experimental neuroscientists from CARMEN, the *Metadata of the week* process was criticised as having been unwieldy, but it was also appreciated as a genuine attempt to involve consortium members in agreeing definitions.

[28] Notes from interviews conducted during June and July are reproduced at Part 4 of this study report

8. Minimum information about a neuroscience investigation (MINI)

It was proposed and agreed at the October 2007 consortium meeting that criteria for minimum metadata would be set, and data could not be uploaded to the CAIRN if this minimum were not provided. Drop-down menus would be provided at the online interface to reveal lists of terms from which intending data depositors could select; a glossary of approved terms was already under development.

By the time of the next meeting, in April 2008, a group consisting of CARMEN neuroscientists and neuroinformaticians had devised and posted on the Web details of a module that would identify the minimum information required to report the use of electrophysiology in a neuroscience study.[29] It was described as the first module in the MINI family, in which each module would represent 'the minimum information that should be reported about a data set to facilitate computational access and analysis, to allow a reader to interpret and critically evaluate the processes performed and the conclusions reached, and to support their experimental corroboration'.[30]

8.1 The MINI philosophy

Quite simply, each MINI module provides the researcher with a checklist of essential information that should be provided when uploading data to the CARMEN system. Its purpose is to describe the data being submitted (including its experimental context) in sufficient detail to allow another informed data user to understand and critically evaluate the interpretations made, and the conclusions reached, as well as to support any experimental corroboration of the original research.

It was acknowledged that the technology independent implementation of MINI, in which neither data format nor repository structure are identified, is indebted to the design principles and presentational format of the MAIPE guidelines,[31] which had been published the previous August and would have been known to members of the Metadata Working Group at the time of the October consortium meeting. Importantly, this approach reinforced the CARMEN WP0 team's commitment to build not only on existing eScience technologies when designing the system architecture, but also to draw on established methods for signalling relationships between data and metadata.[32] [33] The MINI family of guidelines has itself been registered with the MIBBI registry[34] and all modules are currently structured using the FuGE data model.[35]

A further characteristic of the CARMEN approach that has been conspicuous during this study, and which has consistently differentiated it from some other systems implementations, is its focus on practicability - a factor that attracted considerable user approval. In terms of achieving compliance with the MINI checklists, team members were adamant that the process should not in any way create a burden for data submitters, such that it might prohibit the data submission process or deter the use of the CARMEN system.

The reporting requirements for the use of electrophysiology, as defined in the first module, are given in Appendix 1 to this report. They identify a significant degree of detail that must be captured about the data subject, the task or stimulus employed, the protocol(s) and equipment used to record the measurements, and a description of the resulting time-series data. Whilst the

[29] Frank Gibson et al, *Minimum Information about a Neuroscience Investigation (MINI): Electrophysiology*, http://precedings.nature.com/documents/1720/version/1/html

[30] Ibid

[31] Chris Taylor et al, *The Minimum information About a Proteomics Experiment (MIAPE)*, Nature Biotechnology, August 2007

[32] Paul Watson et al, *The CARMEN Neuroscience Server*, http://www.allhands.org.uk/2007/

[33] Examples of complementary services that were researched by CARMEN include the MIBBI project (http://www.nature.com/nbt/journal/v26/n8/full/nbt.1411.html) and the data integration strategy of the Open Biomedical Ontologies (OBO) consortium, described at http://www.nature.com/nbt/journal/v25/n11/abs/nbt1346.html

[34] http://mibbi.sourceforge.net/

[35] http://fuge.sourceforge.net/

list was regarded as a minimum, its length is quite apparent, although anything less was expected not to facilitate the effective interpretation and assessment of electrophysiology data and metadata uploaded to CARMEN (or any other repository). This view had been propounded not merely by the computational experts in the MINI group, but also the contributing neuroscientists.

8.2 MINI in practice

As predicted at the October 2007 consortium meeting, the CARMEN community were to be asked to test the metadata schema that was about to be built, and to advise which terms were thought to be missing. Hence, it was believed that CARMEN's metadata schema for experimental neurophysiology would be based directly upon discipline knowledge and the practical needs of those who would be submitting data. This user-orientated approach was underwritten by a new commitment given at the meeting to use student effort in an investigation of what researchers currently do with their data, and to identify whether there were further needs to be addressed.

Unfortunately, at the April 2008 consortium meeting, in the discussion that developed from a demonstration of the CARMEN prototype, the project team came under fire for the length and complexity of the process for metadata assignment. Further, it was subsequently reported by WP0 to the June project progress meeting that, during the programme of usability testing, which had commenced immediately following the launch of the CARMEN portal in April, the interface for entering metadata was identified as a key failing.

An interview with a key experimental neuroscientist from CARMEN, conducted on 10th June, produced comments on a usability test from the previous week, where the MINI-inspired metadata proforma was criticised as being "excruciatingly tedious", having taken forty minutes to complete for a single file of data. This Web form had "too many fields to complete, many of which are not felt to be necessary or relevant". On the basis of these comments, one might deduce that the MINI approach had been exposed as requiring more than the absolute minimum; certainly there was a call for greater flexibility in what must be completed and, without a change in requirements, it was predicted by this otherwise enthusiastic early adopter that the scheme would not be used.

The concept of developing a standard, however, remained essentially attractive to these practising researchers, as was the promise that the need to input large amounts of metadata at every data upload would decrease as the metadata architecture fills out. The MINI group had already acknowledged in their Web posting that much of the information required may already be stored in an electronic format, or exportable from instrumentation, and they anticipated further automation of the assignment process.

During another interview on 10th June, the MINI approach to assigning metadata was approved on the grounds that it forced consistency, which was judged a significant step forward; and its basis as a minimum requirement remained sound, having been derived from the observation of laboratory practices. Whilst acknowledging that the process failed to avoid being a burden, at least in its initial implementation, this interviewee was prepared to support the application of MINI as a crucial solution for naming and identifying neurophysiological data. These are infamously heterogeneous, as all parameters are liable to change (e.g. spike trains will be run at all kinds of rates), and traditionally have proven resistant to the definition of a consistent set of metadata showing how the data have been gathered and what is being represented by the data recorded.

The predicted simplification of the metadata ingest process, which would arise once the body of metadata matures, was for this second interviewee enough to offset current difficulties; as would the promised insertion of Help links, to explain the structure of the metadata hierarchy, thereby enabling one to see how the layers of descriptive information relate. Indeed, when referring to the CARMEN metadata architecture needing greater flexibility, in this case it was to allow for the ***addition*** of terms that will cater for the heterogeneity between laboratory experiments.

Overall, these somewhat opposing views tend to sustain the concept of the MINI approach. For the Metadata Working Group they probably represent success, since they have provided a logical and sustainable structure where previously there was none. The politics of persuasion and inclusion that lay behind the *Metadata of the week* initiative will have paid some dividends here, but the goodwill generated will fail if the requirement to complete all fields in the metadata submission form does not reduce and it is not replaced by a slicker automatic process for the population of fields.

Nevertheless, during a series of investigative reviews undertaken by project management between July and September 2008, a number of consortium members continued to express uncertainty as to whether the minimum reporting frameworks were effective in capturing their requirements. In the light of the heterogeneity of the user groups, it appears that the frameworks may be 'too generic', making it difficult for users to understand how they should apply them to specific use scenarios. Fortunately, at this point in the project, and in consideration of project scale, it is still possible to enable closer working relationships between the personnel responsible for the development of these frameworks and the individual user cohorts, in order to achieve implementation through co-operation. In a larger and even more heterogeneous service environment such an approach may not be as feasible.

In an email dated 3rd July[36], the Technical Moderator confirmed that the MINI document should be regarded as simply a statement or list of required terms. As he pointed out, the information it provides has to be modelled, and the CARMEN team has chosen the standard FuGE[37] data model to shape the information provided by the MINI framework to represent relationships, cardinalities and referencing. The team wishes also to curate the values used to describe the experiment with the model (i.e. the text the user enters), using ontologies as a mechanism for control and for defining the values the user enters. These would appear in drop-downs in the Web form interface. For this purpose it is planned that CARMEN will contribute to and re-use the OBI ontology for scientific experiments.[38]

In summary, he defined the methodology as:

- Define the information to be modelled (requirements / MINI doc) ->
- Model the required information ->
- Control and define the descriptive values (ontologies).

'The result of us doing this and then users applying this to a piece of data is, I suppose, the "CARMEN digital curation lifecycle".'[39]

The testimony of this observation was its confirmation that the project was proceeding well according to its stated requirements; yet the linear nature of the lifecycle it depicts also underlines that CARMEN is not, first and foremost, a digital curation initiative.

[36] Email from Frank Gibson to Graham Pryor, Thur, 3 Jul 2008 14:56:20 +0100 [03/07/2008 14:56:20 BST]

[37] http://fuge.sourceforge.net/

[38] http://obi.sourceforge.net/

[39] Frank Gibson, Ibid

Part 3 – Engagement with data curation issues, drivers and constraints

9. Analysis – April 2008

The data curation aspects of CARMEN were described in a presentation given by the author of this report to the April 2008 consortium meeting, which set project activities in the context of the DCC's curation lifecycle model.[40] They fell into two categories:

Positive	Negative
Data curation provides an implicit rationale for a (data orientated) community dispersed geographically and temporally	CARMEN researchers as a body are not familiar with the concept and techniques of digital curation
CARMEN has the basis for a solid metadata architecture (using FuGE) that is also attached tc a global development model (OBI)	IPR and data ownership issues have not yet been resolved
The facility for enabling user annotation is directly supportive of data quality, re-use and added value	A vision of open data for CARMEN requires a licensing mechanism
Provision is made for data migration, incorporating the use of 'virtual machines'[41]	Integration of several products and the use of immature technologies carry a risk, where there are only limited opportunities for the project to influence them[42]
Raw data remains static; metadata evolves. Data integrity is preserved; the science has room to develop	Are there appropriate processes for data appraisal, selection and destruction in CARMEN for it to be described as containing a trusted body of data?
The emerging policy for access management, based on the concept of groups, expands on the traditional culture of data depositors whilst also reflecting the new culture of social networking	

The list of negative aspects in effect reflected the status of a work in progress, where none of the issues described would be expected to persist long term. Nonetheless, in terms of awareness of data curation issues, matters had not developed significantly since a review meeting with WP0 members in February 2008, when it was observed that researchers in the consortium were unaware of the data curation issues and had not engaged with these aspects of the project. At that time the focus of the development team was to "build what the users want; they need to see something that can be used. Until then they will be undecided about several project issues and will continue to be fluid in their perception of the project, its aims and deliverables".[43] The driving force was still coming from the system builders, although they recognised an urgent need to make the system live, following which it was anticipated the neuroscientists would do more themselves to drive the project. As a corollary, it was also anticipated that engagement with data curation issues would then follow.

Yet whilst by early April the technical infrastructure seemed poised to deliver the key project aims of unlocking data, the introduction of a shared language, and with shared experiments promising to

[40] http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf

[41] The migration of data/file formats is expected to be tackled by the use of 'virtual machines' in a Linux environment, enabling the recreation of previous versions of a specific software product

[42] CARMEN is not a completely new system but an integration of several existing eScience technologies. Products being used tend to be relatively new and are not always mature; if they do not work properly, or they work in a manner that does not meet requirements, the CARMEN team have limited means of forcing changes.

[43] Frank Gibson, quoted from CARMEN study reflective meeting, Gibson/Knowles/Pryor, Newcastle, 7th February 2008

optimise the benefits of eScience, it was increasingly evident that project success would be dependent on changing the underlying research culture, which would have much to do with replacing the old trust processes based on closed networks and 'tribal' relationships. Acceptance of the full ramifications from the CARMEN rationale was not yet a given in the universal sense, which could also be seen in the context of data curation.

10. Data security and data sharing – a paradox

A view held amongst the CARMEN project membership is that data security is critically important to experimenters. It is also firmly believed that they do not wish to be merely anonymous contributors of data but expect to be directly involved in any further analysis of their data sets, and that this expectation will be supported.

Together with metadata, the rationale for ensuring data and system security was of particular concern to the October 2007 consortium meeting, and in the ensuing weeks it became a topic that generated considerable debate.[44]

10.1. Access management

It was reported to the October meeting that a Web 2.0 style of approach to access management was being considered for adoption, with *Facebook* having been taken as a model for the creation of groups/collaborators. In that setting, all members of the CARMEN community could upload data, but only nominated individuals (e.g. the collaborators in a specific programme of research) would be able to see those data prior to the publication of papers based upon them. Consequently, there would be a need to define which groups should be given access to any given data set. One further option being explored was to identify a 'collaboration manager' with responsibility for controlling access by the broader community.

By August 2008, this ethos had undergone quite significant changes, although these have not yet been documented and approved. The most recent view within WP0 is that data owners (i.e. the scientists who have produced and/or uploaded the data) should have complete control over how their data is shared. This construal of 'data owner' may itself be challenged, since IPR could rest with employing institutions, for example, but it is being promoted as a pragmatic interpretation. There would, however, be an expectation that data will be released to all CARMEN users after publication. The issue here is the definition of 'publication', since if it refers to the publication of papers derived from the data, there could be very long delays before data are released. Meanwhile, the option of employing financial or other incentives to improve the probability of early data release is being explored.

Another interesting notion that has emerged since the April 2008 consortium meeting is the position referred to by some of the WP0 team as the 'spiritual' ownership of data, which meaning is assumed to incorporate moral rights and responsibilities. Whilst data ownership is difficult to define, it is thought that it could be managed by allowing data uploaders to specify and invoke contacts and/or negotiation with other CARMEN users (or individuals outside CARMEN) who may wish to assert ownership, thereby allowing claims to be registered without placing rigid boundaries around ownership. For those who are having to put systems in place this may be seen as a welcome escape from the prescriptive formality of defining ownership; but it could have the opposite result by inviting a broader set of highly competitive individuals to wrangle over who had the predominant set of rights or authorship! If this notion is pursued the outcome will be watched with interest.

At login, users will be able to see their own data and data for which they have been granted access. Permissions governing access by others would be assigned at the point when data are being uploaded, with the CARMEN system able to support defined groups. If someone outside a permissions group wishes to see data an email is generated and permission is sought from the data owner (depositor). It has to be remarked that, in practice, this is expected by some members of the WP0 team to prove onerous for data owners and could, over time, bring about a greater acceptance amongst the researchers of providing more open data.

The concept of groups was expected to be intimidating to some users, even though the Web 2.0/FoaF[45] type of approach allowed researchers to select 'None' when identifying individuals

[44] Further recorded views from some of CARMEN's experimental neuroscientists on the topics of data security, validation and sharing are given in Part 4 of this study report

[45] FoaF (Friend of a Friend) is a term used by online social network services for an ontology describing persons, their activities and their relations to other people and objects

who could see their data. Yet it was a positive attempt to scale the recognised layers of trust within the domain:

> **researcher/supervisor (trust); collocated researchers (may be trust); networking (new trust).**

A particular problem identified at the October meeting concerned the provision of access to derived data. If the collaboration manager wished to grant access to an individual researcher from outside the group, would it not be necessary to obtain the agreement of the experimenters who produced the raw data from which the derived data was produced? Given the mobility of research group members this could prove to be difficult. No resolution to this question was forthcoming and it was referred back to advisers within the DCC[46]. However, whatever the legal position, the protection of data and security of access were agreed to be the crucial and over-riding factors; from the mood of the meeting it was evident that the majority of neuroscientists present were extremely cautious about allowing access to their research data, and they were unlikely to concede much movement on this point.

WP0 members advised the meeting that, in practical terms, access criteria would be met through the application of a set of predetermined policies, with data depositors required to select from a predefined range of categories (i.e. from open access to private). It was originally envisaged that the default upload would be for open access (i.e. anyone in the consortium can see the data), which means that depositors would be forced when making their selection to think about the implications of restricting access or enabling open access. This practice was expected to prove itself as an instrument of change. However, further research into the requirements of the user base has challenged this assumption.

Further, the team acknowledged that there will be a need to protect services (e.g. algorithms) as well as data, at least until publication has been achieved. Referring back to the metadata debate, it was pointed out that additional metadata would be required to advise the utility of tools or models provided with the data as services, since whilst some have been found to be effective and work reliably, others are known to be poorly constructed and can have inadequate user interfaces, especially for anyone who is unfamiliar with the data. Warnings about known bugs and other idiosyncrasies must therefore also be given.

10.2. Data sharing

The data sharing policies being produced at this time by the UK Research Councils (RCs) and other funding agencies provided a parallel landscape to the CARMEN aim for shared experimentation. This new policy environment was not, however, a key driver for the CARMEN rationale. A view expressed on more than one occasion by CARMEN members was that, with respect to emerging national data policies, one has to bear in mind that members of the RCs are service providers rather than active researchers, and as such are not always wholly cognisant of the prevailing pressures and practices within the research culture.

Despite the provision of data management and/or sharing policies by the majority of RCs, their position was deemed to have remained unclear, since they tended to avoid intervention at any level of detail, usually requiring only that data from research they have funded is appropriately shared. With the very real possibility of commercial interest in neuroscience research data, as well as from medical charities, the more insistent challenge for CARMEN was to explore options for the licensing of data. This would be politically difficult for an organisation with no legal

[46] This was the beginning of a continuing dialogue with the DCC's Legal Services Associate who, in this instance, referred the project to the JISCLegal resource on ISP liability, where there were potential analogies for CARMEN. It was also advised that under the EU's e-commerce directive, if CARMEN managers are providing a data hosting service without having direct knowledge of the data content, they would be covered against liability by having a notice and takedown policy (i.e. stating "in the event that…we are not liable"). The topic of ownership responsibilities for derived data was, however, known to be a current and open subject of research.

standing, since whilst it could be argued technically that data deposited in CARMEN are owned by the consortium, individual contributing members would consider that they retained ownership of their data and that CARMEN is simply a neutral host.

This view was reiterated during interviews with experimental neuroscientists in June 2008, when it was claimed that IPR will be retained by the data producer (and his/her institution). A practical solution to the unresolved question relating to the 'chain of use' (i.e. where permission to use one's data has been given and this results in the production of new data which are then accessed by a third party) was believed to be a requirement for the producer/custodian of the original data to be contacted to confirm access, which essentially reflected the old culture of one-to-one contact, although the use of data licences was seen to be potentially beneficial as a mechanism for clarifying chain of use issues. Unfortunately, the vision of open data being managed by some kind of licence, which could be granted or deployed at data upload, has yet to be debated and resolved by the CARMEN community, and as yet there is no agreed mechanism in place.

These interviews served to confirm the WP0 team's assessment of data sharing as a contentious proposition, with a probable majority of researchers perceiving a real need to protect their data from unscrupulous data scavengers. Indeed, the highly competitive nature of the domain does not fit easily with the precepts of eScience. Hence, consortium members would be unlikely to consent to complete open access and would expect to exercise control indefinitely, since data will retain value for future use and exploitation. The risk was even extended to the provision of open access to metadata, a fear that had already been voiced during the April 2008 consortium meeting, as competitors viewing precise metadata may be able to usurp IPR and gain an advantage, particularly where a new research method is being developed.

The considerable investment of money and time in a single neuroscience investigation, and the difficulty of arranging individual experiments and of repeating them, was seen as characteristic of the domain and the basis of an argument against uncontrolled data sharing. One senior researcher explained that a single series of experiments with primates typically takes three years to complete, including the training and preparation of the subject animal, with a consequent and considerable intellectual and monetary cost. It is common, therefore, that the data produced from such experiments are regarded very much as private to the researcher, and there is little willingness amongst them even to share their metadata until after the data have been used to produce published results.

A main tactic employed by CARMEN's project managers, as a counter to this traditional reluctance to share, has been to concentrate on the involvement of early adopters who want to 'get the data out quickly' and prove the system works, and who might by example encourage a sense of comfort amongst others; but even these have retained some serious reservations.[47]

Another side to this issue was described by the Project Manager.[48] In terms of project expectations, he perceived that one of the researchers' principal views of CARMEN is of a new vehicle for enhancing their ability to publish. Traditionally, because of the cost and difficulty of conducting the research and producing the data, they have tended to horde data and trickle out a series of publications, following a kind of subsistence model. If one identifies data curation as 1) involving the preservation of data, 2) enabling optimised access to it and 3) adding value, it is acknowledged by the WP0 development team that the curation ethos is an underlying motivation for the project rather than an overt one. However, with neuroscience researchers spread both geographically and over time, data curation may be argued to provide an unacknowledged yet essential rationale for eventual change with respect to data sharing.

This rationale is, however, proving slow to take effect. In the investigative reviews conducted by project management between July and September 2008, some members remained very reluctant to disclose collaborations with individuals external to the consortium, which reluctance could

[47] See Part 4 to this report

[48] Alastair Knowles, CARMEN study reflective meeting, Gibson/Knowles/Pryor, Newcastle, 7th February 2008

extend even to a disinclination for making introductions between established collaborators and prospective collaborators. As already indicated, the culture for trust and collaboration in neurophysiology is predicated on lengthy 'courtships' before relationships will mature to the point at which ownership of pre-publication resources (e.g. data, code) is routinely exchanged. This attitude continues to challenge the short term viability of CARMEN not only as a virtual laboratory but as a mechanism for pre-publication data sharing. Nevertheless, a significant number of members confirmed that the emerging policies of both funding agencies and journals, concerning the dissemination of post-publication data, are having to be recognised as a powerful driver for change.

That said, until a cultural shift is achieved, the short to medium term strategy for CARMEN will have to be based upon post-publication data sharing, and it is being recommended that CARMEN should be active in engaging journals to promote this approach[49]. At the same time, incentives are to be explored with a view to rewarding those wishing to disseminate data early in the publication cycle. This would provide a financial advantage, as it should encourage new users to subscribe to CARMEN in order to access public resources.

In summary: experimental neuroscience data are not widely shared prior to publication mainly because they have high perceived value. This perceived value is a consequence of highly individualised research that is often unique to the environment in which it is conducted. As observed by the Project Manager: 'the domain norms underpinning sharing are directly related to the cost of acquisition and are outside the control of individual actors, including CARMEN. However, long term development of high throughput acquisition platforms may transform the domain norms in favour of CARMEN.'[50]

10.3. Validation and quality

The October consortium meeting provided an opportunity to seek members' views on measures required to address the question of data quality for data that will be 'published' through CARMEN - i.e. what will stand instead of the peer review process that is rigorously applied to research articles? The view supplied by the WP0 team, and corroborated in plenary, was that perceptions of data quality will be dependent upon the established reputations of the data depositors and their laboratories or institutions. Data from unfamiliar sources would automatically be treated with appropriate caution.

A more emphatic answer was provided at a meeting in February 2008, when the Technical Moderator indicated that the validation of quality (of content) is not an issue for the project: "CARMEN is not a peer review system. Its mission is to make data available and to describe the data, but not to exercise any judgement on it".[51] However, annotation does represent a key feature of the metadata description, where researchers can enter individual judgements on the quality of specific data sets contained in CARMEN (i.e. how easy they are to use and how useful, whether they set a gold standard, etc.), which is a unique means of exposing researchers to their peers. This aspect of the system has still not been fully recognised by the CARMEN membership and may in due course be a cause of some apprehension amongst researchers, although it should lead to improvements in the quality of the data uploaded when it is known that poor data could be publicly exposed to criticism.

As would be normal practice, in terms of system services it is the responsibility of the WP0 development team to provide mechanisms that will prevent the upload of viruses, DNS attacks, etc. Yet this technical rigour sits uneasily with the content of the CARMEN consortium agreement, which does not qualify how members use the system or specify their responsibilities with respect to data sanitation.

Issues around misuse have not yet been adequately addressed. Following the beta release in April 2008 it was expected that the WP0 team would be able to monitor how CARMEN is being used;

[49] A strategy for engaging with discipline journals is described in more detail in Part 12

[50] Alastair Knowles, unpublished report on discussions with project partners, September 2008

[51] Frank Gibson, quoted from CARMEN study reflective meeting, Gibson/Knowles/Pryor, Newcastle, 7th February 2008

but pragmatically, in the context of the neuroscience culture, the Project Manager has agreed the emphasis has to be on growing the user base in a context of trust, facilitated by the social networking approach that has been adopted.

11. Sustainability

There are no guarantees that the pilot CARMEN system will be maintained beyond the end of its project phase. Notably, and on a purely practical level, the servers used to enable the CAIRN have been bought with project funds and currently there is no funded replacement plan; nor indeed has any provision been made to expand storage capabilities beyond the lifetime of the project.

Nonetheless, confidence in CARMEN as a substantive data management platform has already been implied. By early 2008, CARMEN was being quoted extensively by consortium members in their grant applications, where the infrastructure represented by CARMEN was proving beneficial in helping meet the new data sharing imperatives imposed by the research councils.

As an enabling force, CARMEN has been demonstrated to have stimulated cross-discipline collaborations/funding, with particular success in respect of Newcastle-based projects[52], in addition to which the opportunity for collaborations with similar platform providers in the USA and Japan have been (and continue to be) explored. This promise of a broadening user base suggests one option to demonstrate the project's success: a 'map of collaborations' would illustrate the increased granularity of the CARMEN community and would underwrite the project/system as a credible platform for research and data sharing.

With the potential user base running into thousands, the implications for data curation by a sustained and expanded CARMEN service remain a considerable challenge, since this user base continues to be composed of hundreds of 'cottage industries'. This makes it no less useful to attempt to share/compare data, but any designs that treat the curation requirements of the user base as homogeneous may inhibit uptake (when, as a consequence of efforts to accommodate the spectrum of identities, the archiving and metadata frameworks are made overly generic and restrictive). This lesson had already been learned during the implementation of the MINI framework.[53]

In terms of the identity that CARMEN might develop, the series of investigative reviews with CARMEN members that was conducted between July and September 2008 confirmed that user requirements remain data orientated. The following analysis has been adapted from the report on those reviews. [54]

> A minority wish to use the system to access clusters, in order to perform long-running/intensive computations, but the majority cite access to data, the ability to share data with collaborators (e.g. analysts), and the ability to discover and form collaborations as their primary motivations for engaging with the CARMEN system. Indeed, analysts have an inherent desire to disseminate code, as this leads to new applications and publications, and they have identified CARMEN as a platform for this purpose.
>
> Recently, a number of the CARMEN partners have been provided with access to very large processing clusters provided by their institutions, a change that has been enabled by the falling cost of processors, but CARMEN has not 'scaled up' in the same way. Looking ahead, one option for CARMEN has always been the provision of data processing services to other organisations, as a third party model offers economies of scale, but with the compute landscape having changed, there is more onus on CARMEN to enable the sharing of data and algorithms, and less to provide access to actual compute power.
>
> The technology landscape will be key to the future of CARMEN. Institutional clusters typically support all domains and an outsourcing solution for one specific domain does not represent good value at the moment. In respect of compute provision, it has been argued that this leaves CARMEN with three possible modes of operation:
>
> 1. CARMEN to be offered as a domain generic platform with cluster outsourcing, supporting user deployed services. Hosting and support to be provided remotely.

[52] Alastair Knowles, unpublished report on discussions with project partners, September 2008, Appendix 4
[53] This report, Part 8.2
[54] Alastair Knowles, unpublished report on discussions with project partners, September 2008

2. CARMEN to be licensed as a peer-to-peer deployment platform for generic or specific domain use. Hosting and support to be undertaken by the institution.
3. CARMEN to be offered as a domain-specific platform, with limited analysis functions allowing data/services to be evaluated. Users to download or ship resources to their local processing hardware to perform more complicated computations.

The first option is judged to offer economies of scale by aggregating the cost of hardware, facilities, maintenance and support. However, such a solution requires a platform infrastructure that is generic, scalable and self-supporting, as well as remote user support, none of which are deemed to be realistic aspirations in the short term.

The second option allows institutions to deploy their own CARMEN peers. This allows maintenance and support to be devolved to institutions. However, it places significant onus on the maturity of the infrastructure, as institutions will not bear the risk of deploying unstable products on their own hardware. Similarly, the deployment of unknown data/services from external peers will be perceived as a security risk, prohibiting sharing within CARMEN.

Option three proposes reduced functionality, meeting core requirements at low cost. Support for user deployed services may become critical at a later stage.

The project manager's recommendation was that option 3 provided the best framework for continuation, one in which funding may be sought in conjunction with continued research into web service technologies. There is potential synergy here with the interim recommendations from the UK research data service feasibility study[55] (UKRDS), which is expected significantly to influence national strategies for digital research data services. In their interim report to the UKRDS Steering Committee[56], published on 7th July 2008, Serco Consulting argue that concentrating on a hybrid/umbrella solution offers the best means of fostering the coherent development of technology and standards, as well as a vehicle for assisting existing service providers in developing and marketing their services and facilities. It is possible that CARMEN could function as a discrete service beneath such an umbrella, retaining its integrity as a domain specific solution.

This approach will not, however, be of value to CARMEN (or similar domain initiatives) without substantive restructuring in the provision of funding for informatics-related activities in the UK. A business model that will sustain CARMEN beyond its project life requires recurrent funding, not only for capital replacement but also for essential and manpower-expensive infrastructure services with functions such as data hosting, data preservation, and access control. For a single project such as CARMEN these produce a large overhead that threatens to make the data sharing platform unsustainable. It is not clear whether the UKRDS has addressed the cost of duplicating these costs across the domains to be brought under its umbrella.

Serco Consulting do, however, observe that metadata standardisation may be intractable for many domains, which important analysis confirms a view emerging amongst CARMEN project management that a far more pragmatic understanding of where standardisation may add value, and where it may actually impede or prevent data sharing, has to be based on the maturity of the individual domain.

Reflecting on the UKRDS interim report and its implications for neuroinformatics, the CARMEN Project Manager describes this view as coalescing into an 'option for one (or a small number of) very large and centrally coordinated data sharing platforms providing high quality generic services at low cost. As a neuroinformatics project, [CARMEN] would like to build [a] *boutique* [of] neuroscience applications on top of one or more of these generic platforms. This represents sustainability. The platform resources represent viable enterprises as they can serve thousands of users at low (or potentially no) direct cost. Projects like CARMEN are more

[55] A study into the feasibility and costs of developing and maintaining a national shared digital research data service for the UK Higher Education sector - http://www.ukrds.ac.uk/
[56] Publicly available at the above URL

sustainable as they can conduct their science at an ongoing cost that is more palatable to public research funders.'[57]

Meanwhile, the CARMEN project team is faced with a seemingly intractable impasse. From the mid-2008 series of discussions with partners it is apparent that most of their end users are scratch coding (i.e. developing simple scripts in environments such as MatLab and R, which are not broadly intended for re-use). This is leading towards a stark disparity with the intended CARMEN user environment, which would be supported by services, since Web services normally provide a means of publishing more mature applications. There is a perception within the project that other domains, in which researchers might possess better programming skills and are accustomed to sharing applications, would find Web services more suitable; nonetheless, they are pushing ahead with the services agenda, but refocusing on generic algorithms that all of their users need to use, intending to provide a fairly high level of support for service wrapping. Of course, the process of trying to nail down generic algorithms is fraught with frustration, and mirrors the MINI experience. There are always so many special cases to be accommodated.

In Part 10.2, reference was made to the active engagement of journals in the promotion of post-publication data sharing. It is worth noting at this point that such links offer a further detail for the developing CARMEN sustainability strategy that could impact on the data curation model for experimental neuroscience data. Here, the benefits from substantiating publications with source data may be used to develop mutually advantageous relations with journal societies, including the provision of a 'Publish in CARMEN' service for authors.[58]

Despite the inclusion of commercial partners throughout the project phase, the preferred platform for CARMEN's transition into a sustainable entity is the creation of a not-for-profit organisation. The continuing internal argument over the dynamics of a sustainable organisation is, however, illustrated by the Project Manager's justification of a standalone entity, which he suggests would be free from potential conflicts of interest that arise between rival host institutions in a consortium.[59] The adverse fall-out from such conflicts is viewed as the possible reluctance by third parties to invest in CARMEN services. Interestingly, he also notes that major stakeholders such as pharmaceutical companies should be included in this discussion. This introduces a new tension into the concept of providing a not-for-profit service to a data-rich environment.

At the conclusion of this longitudinal study of CARMEN (but not the conclusion of the CARMEN project), the project team has rehearsed its sustainability strategy as a series of steps, with the final goal of not being directly dependent upon public research grants. The anticipated next step is a platform grant to harden the pilot CARMEN infrastructure into something that can support a scalable service, and overtures to the BBSRC[60], Wellcome Trust, TSB[61] and now the UKRDS are to be orchestrated in the foreseeable future. Each of these bodies has made a recent and explicit investment in data sharing strategies, and it is the data management value of the CARMEN solution that will be argued as a principal claim for longevity.

[57] Email from Alastair Knowles to Graham Pryor, Mon, 22 Sep 2008 10:40:44 +0100 [10:40:44 BST]

[58] On submitting a paper for publication in a journal, authors would be asked if they would like to publish source data in CARMEN, and a referral mechanism to CARMEN services would be supplied; the journal society and author(s) would then exploit this service to promote publications that reference source data, which would generate greater interest

[59] Alastair Knowles, unpublished report on discussions with project partners, September 2008

[60] Biotechnology and Biological Sciences Research Council

[61] Technology Strategy Board

Part 4 – Concluding perceptions

12. The CARMEN lifecycle

In his summary of the methodology for CARMEN metadata, the Technical Moderator referred to a particular interpretation of the "CARMEN digital curation lifecycle".[62] That was to take a corporate CARMEN view, which was reiterated by another senior member of the WP0 team, when discussing the DCC curation lifecycle model:

> **"Well, we don't do appraisal and selection because that's the user's job. Likewise transform. And we have no preservation actions at the moment, because we neither know how to do this nor understand what is required.**
> **All the rest, yes."[63]**

Whilst from the perspective of a data curation purist this analysis may disappoint, the position taken here demonstrates the focus of the CARMEN team upon their principal objective. It also reveals something of the nature of the CARMEN community and of project dynamics.

These issues were further explored during a final meeting with the Project Manager and the Technical Moderator, in Newcastle, on 5th August 2008, when they were invited to reflect upon the progress made by the project, particularly in the context of dealing with data, and how its complexion had evolved. The main themes from that meeting are reproduced here as a personal analysis of the CARMEN organisation, its conduct, method and progress. The headings (underlined) were supplied in advance as prompts. The views expressed here are those of the two interviewees, not the author of this report.

Some of the material reproduced here has already been incorporated into the body of this study report.

12.1. Reflective analysis, 5th August 2008

Project aims and objectives – how have they changed?

There has been an increase in emphasis on the priority to deliver software to the users. This has required a greater degree of engagement with third parties and a significant level of usability testing. The project remains very much 'on spec' and user driven, as was planned. Requirements remain datacentric, from the points of view of both the data producers and the code developers.

In terms of project progress, the nature of a distributed development programme has been shown adversely to impact pace and communication. Strong central management is essential to articulate project direction; but it was a year before a project manager was appointed, which was not the easiest of positions to be in with respect to applying direction and leadership.

Having two development teams, at two geographically distant locations and each of them familiar with different methods and cultures, has also introduced some inevitable misunderstandings and delays. The development of a software product that will support actual services is not normal business for an academic department and has required the acquisition of new skills and aptitudes. For the developers, working on such a large, multi-site development has represented a new challenge.

The composition of a mixed project team (computer science, bioscience, etc.) also brings together competing aspirations and values. Some team members are happy to launch a working system, others would prefer to wait until the system has been fine tuned. This situation has led to frustrations amongst consortium members eager to load data.

Evidence of influence by data curation considerations on the shape and direction of the programme?

Usability testing has exposed the need for a mechanism that will allow users to manage their files on an individual basis (e.g. a filesystem). Users have amassed a large number of discrete data files which need some means of structured management ("I do this now…I want to see…").

[62] Quoted on page 21 of this report

[63] Email from Phillip Lord to Graham Pryor, Thur, 3 Jul 2008 14:12:26 +0100 [03/07/2008 14:12:26 BST]

If one considers that data curation comprises preservation, access and adding value, the processes for annotation and data integration are what CARMEN is enabling. However, the effectiveness of CARMEN in this area will not be known until the system is fully live, with users actively uploading data, making annotations and undertaking queries.

The introduction of a comprehensive and appropriate ontology (currently under construction), which makes data usable across projects, will impact considerably on the outward-facing complexion of CARMEN.

Changes in awareness/requirements of consortium members with respect to data curation aspects of the CARMEN project?

There is still some ignorance of the issues and demands surrounding good data management, but in general there is a good understanding of the need for an effective common language (ontology, metadata, etc.).

Mandatory sharing/publication of data, as required increasingly under the research councils' funding strategies, is creating a new focus amongst researchers. Some within the CARMEN community have already experienced increased citations as a consequence of data sharing.

It will be difficult for CARMEN users properly to understand curation issues until the system is live and they are confronted by them 'in the flesh'.

Since the project was conceived, how core have data curation issues become as drivers or constraints for the project?

Data has been collected over very long periods and, whilst within the domain there is a desire properly to organise and manage it, as well as to share it appropriately, the scale of the data legacy has militated against these aims.

CARMEN has shown that curation and sharing is now an achievable possibility, and is responsible for driving a step change in attitudes and expectations for data that are new to the domain.

Researchers have managed to survive with limited approaches to data management whilst working in a one-lab environment, but when the CARMEN system has grown into a large community the need for structured data will become more conspicuous. In such a community it would be too onerous for researchers to handle a multitude of individual enquiries for data access or to keep notes and annotations on paper.

Status of metadata requirements and solutions

The MINI approach that has been developed is language-independent. It is basically a list of statements that define requirements. Complaints that the scale of the 'list' is burdensome were defused when users were asked to remove those elements regarded as unnecessary or not required – and none were removed. However, the project team anticipates a natural trimming down after the tool is more widely used and practical working requirements are established.

The data model is in place and the ontology is under construction (a translation of the content of MINI). Users will be able to add terms through submission via a Web form, using a drop-down box.

The project team is aware of outstanding questions relating to workflow, particularly concerning the appearance of individual elements in their correct sequence – for example, data may be uploaded without metadata, or if accompanied by metadata may require action on permissions, etc. As its name implies, MINI represents minimum requirements, so one must always allow for the provision of additional terms and to include the consequences in the workflow pattern.

In terms of input, templates have been provided. One selects either data or services, which can then initiate auto-population.

Arrangements for validation and security – current and predicted (i.e. original consortium v global community). Is there a difference?

The consortium's responsibility is to provide an infrastructure, within which methods for annotation are enabled and security assured. Beyond that remit it has no responsibility for making judgements about the data deposited, which responsibility remains with the individual data producers and/or curators. Such an approach reflects Web culture.

The data in CARMEN is structured but there always remains a risk that a piece of malicious software code could be uploaded. The same applies to most Web services.

Quality management governing data selection and deposit – is this/will it become an issue?

The only real issue is that of providing sufficient disk space.

There is an option for separating data into individual networks, as is the practice with *Facebook*.

User knowledge/training – the demonstrable usability of CARMEN; the extent of skills acquisition and coaching required?

Researchers are resilient and traditionally self-reliant, with the majority used to working things out for themselves. CARMEN may be very different to the environment with which they are familiar, which could lead to slow uptake, but they tend to be persistent when acquiring new skills.

The true situation will become clearer when the system is live and users are attempting to use it on a day-to-day basis. Currently only early adopters are providing a measure of user capability. Remembering that the objectives of CARMEN are of a *pilot* project, the software product is the visible deliverable and has yet to be used in earnest.

Education is being achieved through consortium meetings and simply speaking to individual consortium members at any opportunity. Some more structured training has been provided (on Web Services, workflow, etc.).

Ownership, sharing, IPR, licensing – what are the legal loose ends?

The project team have not yet examined the licensing issues in any great depth.

Data in CARMEN are a recording of signals from neurons and are not copyright; the metadata assigned to data may, however, attract IPR and be open to copyright. If the researcher is engaged in producing a scientific method, access to the metadata could cause disclosure and a breach of IPR, which explains the caution some researchers exercise towards making their metadata openly available. Nonetheless, researchers' data is generally protected in CARMEN by the provision of controls that can limit access to metadata only.

The management of chains of ownership, where a new data set may be derived from privileged access to an original data set, but may not be subject to the same level of access control, has yet to emerge as a 'living' issue. However, it is recognised that one of the project challenges is to provide a suitably sensitive means of managing the ownership of derived data. At the moment the users do this informally (e.g. through gentlemen's agreements) and it is difficult to understand how to abstract/codify this practice.

Sharing data in a more open way has yet to become universally accepted. Researchers in neuroscience are currently working in a very granular fashion and have not made the transition to a broader perspective. The CARMEN model will force the issue, since unless users agree to use CARMEN as a means of sharing their data it will be nothing more than a mechanism for providing free disk space.

Looking ahead, the project team recognises the need for some form of licensing, which would govern not so much the right to read data as more particularly the right to use data discovered in CARMEN.

Members of CARMEN have yet to address the proposition that by making data available publicly they are, in effect, publishing it. The traditional notion of publishing persists, where publication is synonymous with the appearance of a paper in a prestigious journal. Yet pressures from the research councils for grant holders to share data with the minimum of delay, and the predicted changes to the RAE (now Research Excellence Framework) are already bringing a new influence to bear. One question they both bring with them is how are researchers to be rewarded for sharing their data, especially when the early loss of exclusive ownership could adversely impact their individual careers. Questions like this pervade the CARMEN consortium in the form of anxiety towards moving too quickly to opening the data for more open access.

Services and sustainability – including a predicted role for intermediation and support (computational); latest best options for post-project longevity; implications from sustaining the pilot as a domain service.

CARMEN has a further two years to run. Beyond this pilot project phase, it would require ongoing software development to accommodate changes and improvements in the science – the need, for example, to handle new experimental methods. This will require the continuation of expert support and maintenance.

The domain is typified by slow growth, reflecting the manner in which researchers combine and start up a research programme, and traditional methods persist. CARMEN is expected to accelerate these aspects of the domain, enabling more rapidly assembled partnerships, and in the future could offer enhancements such as direct, automatic entry of laboratory notes. These are obvious benefits, yet for longer term sustainability it is important to define some metrics for success, based around the continued delivery of a usable system, a user base that has grown/continues to grow, and evidence that CARMEN has directly enabled new science.

To scope its longevity, CARMEN also needs to redefine its customer groups, their long term requirements and the relative cost-benefits of customer-focused services (e.g. Web hosting/low cost/high value, data modelling/high cost/high value).

There is also an outstanding need to develop links with appropriate journals, to offer a data management framework for journals, in which the relevant research councils would endorse researchers using the CARMEN service in conjunction with preferred journal titles.

It may be possible to seek further research council funding on this basis. Already, grant applications now include an option to identify ten year costs for data hosting, which CARMEN should also reflect in its sustainability model.

One option favoured within the development team is to make the CARMEN code open source, hence inviting a broader community than the consortium to contribute to its development.

13. The neuroscientist's view of CARMEN

The following text is a consolidation of notes from interviews conducted with experimental neuroscientists in the CARMEN consortium, based at Newcastle and Cambridge, during June and July 2008. They have been grouped under a set of common headings that were used as interview prompts, which provides a means of concealing individual identities. The interviewees selected are all members of CARMEN science work packages and may be considered to be enthusiastic early adopters.

These notes represent the views of the individual neurophysiologists interviewed, not the author of this report. Some of the opinions recorded here have already been incorporated in the body of this study report.

Carmen project rationale

Originally

The purpose of CARMEN was to build a new environment in which to store, handle and work with data. It was anticipated as a solution to the production of multi-site recordings, which have since become highly prevalent[64].

The purpose of CARMEN from the beginning has been to provide a system in which one can store data, share data, do analysis and share analyses.

When joining the project only a year ago, the purpose of CARMEN was clearly the creation of a collaborative workplace in which one could mine data, initiate new collaborations, develop synergies and, by bringing disparate data and researchers together, undertake new science. This view has not changed.

Formerly one was dependent for collaboration on the transmission of files by email and/or FTP. Working in a globally distributed community meant that one was in constant communication with colleagues elsewhere, and struggling with system inadequacies. A CARMEN type solution was expected to enhance collaborative effort, providing an organised structure for the global exchange of data.

Producing 10gb to 12gb of spike data a day, there is a significant storage and distribution problem, although there is an increased expectation that data will be shared for the progress of this multidisciplinary science.

Currently

Development of analysis tools is a key expectation and a required feature of CARMEN, since there is an urgent need to achieve uniformity across groups, continents, etc. This standardisation in analysis processes will enable researchers' understanding of neural networks to be moved forward considerably.

The opportunity and benefits from standardising processes/tools/methods has become evident.

As a contributor of data, it is apparent that kudos will derive from those who view and use one's data – i.e. via citations. Being able to see other researchers' data may save time in one's own research.

CARMEN will provide a stable environment for raw data. Before CARMEN, data have been stored on a multitude of DVDs, with several copies being made and used. It will be preferable to have an authoritative central repository for the definitive set of raw data, which will be easier to identify and retrieve than when consigned to a pile of DVDs.

[64] It was claimed that these recordings are frequently represented as huge spreadsheets but CARMEN managers advise that, traditionally, most of the data is stored in .txt/.ascii files containing binary sample values (e.g. for continuous signals) or timestamps (e.g. for spike trains)

CARMEN will enforce the provision of proper notes about the data, an improvement on the traditional lab book, making it easier to return in future and access/understand the data.

Retinal research depends on the combination and analysis of data from many different laboratories. Data used to be acquired through request and invitation, which could be laborious. The CARMEN system facilitates data exchange and sharing, providing a new capacity for data exploration that is both wide in terms of reach and inclusivity and deep in complexity and analysis.

CARMEN project methodology and dynamics

The project methodology reflects user requirements and there has been a successful balance of technical staff and scientists, but interaction between the separate work package teams could be improved in order to achieve a clearer view of the bigger picture.

Given the size of the group, the development has gone well. The project manager is recognised as 'doing a particularly good job' in ensuring that the system reflects accurately the actual requirements of the researchers.

Management of the project has been exemplary. "Work package 0 folk are great!" There has been an inevitable computing science approach but the IT team has proven very receptive. Work package 6 has benefited from being a test case, leading to rapid tailoring based on suggestions from the researchers.

The urgency of delivering a solution to the increasing volume of data being produced has caused some frustrations, but it is recognised that CARMEN development cannot be hurried. However, the pace can be irritating: for example, a year ago, Interviewee X was asked to develop a distributed team to work on retinal processes, which was quickly arranged, but the team is still waiting for the CARMEN system to be ready to handle its data.

Being a guinea pig for CARMEN has proved to be a considerable plus, 'having all these computer geeks at my disposal'. Addressing the particular needs of spike sorting, there has been a highly beneficial interaction with the signal detection team at York, who have modified the technical programme to accommodate Newcastle's needs, so that it now works across both environments, with a laborious manual process currently being automated. The usability events (first in Newcastle on 3rd June) are being used to work through issues and fix any problems before the system is released. The project team are 'good at listening'.

CARMEN effectively funds researchers to build new analysis tools, which may be judged as innovation conditioned by the pilot study - a 'clever' outcome of a true eScience package.

One problem (challenge?) has been meshing the competing priorities of the neuroscientists and computing scientists. For example, MatLab is not easy to support and there are serious licensing issues, but MatLab is essential to the science community. The computing scientists attempted to provide an alternative suite of tools but this solution was not acceptable and they were required to deal with the MatLab conundrum. 'No-one will switch to a format or scheme that is different from that with which they are familiar and which is known to be effective.'

Key characteristics of neuronal data

Raw data have infinite value, as interests in the data will change over time, and there is a necessity to look at the data again from a new perspective. For example, one may first want to examine spiking events, then later maybe analyse local field potentials using spectrograms of the raw data. So the data must not be lost.

Recording and analysis tools will change/improve too, so there will be profitable opportunities to return to the original data. The biggest leap is expected in improvements for recording tools. Papers in neurophysiology are cited >40/50 years after they were published, which makes it important to be able to refer to the source data long after they were captured.

Retaining raw data is less expensive than repeating experiments. It is difficult to repeat experiments (changed environment, subjects, location of test 'site'). Recordings themselves are difficult and expensive to make, providing a further reason to preserve data.

Data produced from retinal research can have high medical relevance. In the case of research into epilepsy, data is obtained from patients, with signals recorded from *in situ* human tissue. It is difficult to arrange these data acquisition events/materials and the data is irreplaceable.

The retinal data captured is extremely clean. It is produced from a high sampling rate and demands a large storage capacity. One needs to capture a long sequence of recordings in order to obtain a full and readable 'picture'. It is also expensive to produce, in terms of both equipment and the cost of animal experimentation.

It is costly to keep laboratory mice; the cost of experiments involving primates is considerably higher, where a three year programme can attract £90,000 in animal housing/welfare. Typically, research into a single 'idea' can cost £500,000. This makes the data very valuable.

Some studies use animals as subjects, others use humans and are longitudinal, where there is a natural reluctance by researchers to attempt repeating an experiment or study session, due to the ethics involved in using animals and humans. In the case of longitudinal studies, it is essential to preserve data to enable the mapping of changes in observations over any given period.

Infrastructure for curation

Prior to CARMEN it was a case of storing zillions of videotapes in the laboratory where they were not protected.

Data management is traditionally the responsibility of individual labs. They tend to follow recognisably similar structures but are differentiated by local systems and practices. CARMEN will introduce an environment that absorbs the quintessential elements of local working to produce a centralised, standardised way of working.

A previous neuronal project provides a pre-CARMEN view. It involved 24-hour sampling, with 4 terabytes of data produced per subject animal. Back-ups were made using 1 terabyte hard disks (weighing 22lb each), and a total of 40 disks were kept on a shelf, together with log books describing them and the structure of their content. There were limits on the amount of data that could be kept online, which constrained data use. The data produced were organised according to a standard data format designed in-house; acquisition software was written in-house to meet requirements for capturing data from electrodes at 12kHz (with 7.5 hours of data produced per channel). All the changing, labelling, storing and retrieving of disks was done by hand. CARMEN, by providing an online, shared resource, removes this laborious, physical human role, which was a major bottleneck.

Curation issues

Prior to CARMEN, curation (i.e. the safe storage of data and data sharing) was always a concern, but it was accepted that researchers were just not equipped to deal with it in the most appropriate way, and were left to their own devices on a lab by lab basis. Disk storage in a repository may still be expensive, but it is only a fraction of the cost of videotapes and does at last enable the sharing of data.

Metadata requirements and solutions

In a usability test last week, the metadata proforma was found to be 'excruciatingly tedious', taking 40 minutes to complete for a single file. There are far too many fields to complete, many of which are not felt to be necessary or relevant. Greater flexibility in what must/can be completed will be essential; otherwise it will not be used. The concept of a universal standard is, however, attractive, as is the idea that the need to input large amounts of metadata will decrease as the metadata architecture fills out.

The 'metadata of the week' process introduced by the project team proved unwieldy, but was appreciated as an attempt to involve consortium members in agreeing definitions.

The MINI metadata system forces consistency, which is a step forward. It is based upon the observation of lab practices, with the intention of defining the minimum requirement.

Prior to CARMEN, descriptive but limited file names were used, with identification of the channel number and the environment. This has now been translated to the more comprehensive MINI facility.

The sort of data produced in the lab makes no sense to anyone without metadata describing test conditions, approach, etc., where analytical processes will also differ between labs in considerable detail.

Other sciences have standard databases with which to work – see, for example, what is available for crystallography – whereas neurophysiological data are highly varied, with all parameters liable to change (e.g. spike trains will be run at all kinds of rates). This makes it hard to define a consistent set of metadata that will show how the data have been gathered and what is being represented by the data recorded.

The CARMEN metadata architecture needs greater flexibility to allow for the addition of terms that will cater for the heterogeneity between lab experiments. It also needs the insertion of Help links that will explain the structure of the metadata hierarchy, so that one can see how the layers relate. The reduction to the metadata ingest process, as the body of metadata matures, will be welcomed, as in the initial stages it appears to be a particular burden.

User validation and security

Unfortunately there is a real need to protect one's data from unscrupulous data scavengers, so consortium members are unlikely to consent to complete open access. Neurophysiology is highly competitive. There is a risk even from allowing open access to metadata, as competitors seeing precise metadata may be able to gain an advantage.

The data producer/custodian needs to be able to define user groups, membership of which may be limited to consortium members. This need to control access will be persistent, as the data will retain value for future use and exploitation. However, individual researchers would be happy to receive email requests for access and this function should be built in to the CARMEN system.

There remains a conflict over data sharing. The computing science members of the project want openness; a body of the research contingent does not. From the scientist's perspective, it is not feasible to upload data until findings are published.

Requests for access should always come back to the originating data producer/custodian. People using someone else's data for a publication should make the draft available to the data owner prior to publication. The need for citation (of the original data/data producer) for third or greater iterations in a chain of use will depend on how far the original data have been changed.

Interviewee Y does not generate new data and has no problem with data sharing, although the concerns of other researchers are recognised.

Data selection and quality

The quality of the CARMEN repository/data store is dependent upon trust in the integrity/ability of consortium members. There is an expectation that they will not upload poor data and/or algorithms. Dependency on the reputation of individual labs and/or scientists would be a feasible concept even at a global level, since the neurophysiology community is not huge.

Quality assurance for data being uploaded is based on the reputation of consortium members. "People wouldn't put their crappy data on it" because it would expose their inadequacies. In the same vein, sharing data in CARMEN should improve data quality since it is made open to a broader community of peers.

It would be very difficult to implement central monitoring of the data being uploaded. There are too many science issues to consider and it would need the employment of a considerable team of neuroscientists!

One should be able to post comments about the data, as one can about items for sale on *eBay*, thereby contributing individual evaluations of the data for the attention of other data users.

User knowledge/training

Being a guinea pig from early in the project has been beneficial. The system doesn't look difficult to understand from a user's perspective: the user interface is friendly and all appears very straightforward even for the least IT-savvy individuals. CARMEN has been designed for neuroscientists with the least computing ability: drag and drop functionality is quick, easy and simple.

Manuals will be provided; an online Help system has been requested. No significant training should be necessary except in the use of specific tools provided in CARMEN. This is planned. However, if tools are uploaded that are not supported by CARMEN (and this is possible/acceptable), that will be a problem.

Most of the concepts applied in CARMEN are familiar and accessible. One may need to learn about Web Services, but the project team is generally very supportive and relaxed about providing training as and when it is needed.

Services

CARMEN will sustain versions of software programmes (e.g. MatLab) so that researchers can always use their data without inaccessibility caused by migration drift.

Data ownership

CARMEN will not/cannot become the data owner. IPR will be retained by the data producer (and his/her institution).

There is an unresolved question related to the 'chain of use' – i.e. where permission to use one's data has been given and this results in new data which are then required by a third party. It is probable that the producer/custodian of the original data should be contacted to confirm access. The use of data licences would be beneficial, particularly to clarify chain-of-use issues.

Experiments with primates typically take three years to complete, including the training and preparation of the subject animal, so there is a great intellectual and monetary cost involved. The data are regarded very much as private to the researcher, and there is little willingness even to share metadata until after the data have been used to produce published results.

Data ownership will always remain with the originator or the funding body, never with CARMEN. Some form of licensing will be necessary to govern the use of certain data sets – i.e. to maintain the rights of data producers in different contexts (e.g. citation levels, levels of re-use, etc.).

It is important to make sure the data owner knows who you are. Building relationships matter if one is to 'get behind the data'. Questions such as "what did you find?" and "what's of interest?" will produce the angles and tips that lead to worthwhile further exploration of the data. Openness when seeking to use others' data produces the best results.

The CARMEN pilot

It looks good so far. No negative observations. It seems to be doing what is required by neurophysiologists and 'everyone wants what CARMEN is doing'.

Usability testing has commenced, but with there being only a low level of functionality achieved so far, it is difficult to judge the product/service.

The metadata architecture and assignment process need to be simplified. The metadata interface is "daunting", having too many potential fields for completion. The project really needs to consider how deep it is essential to go when assigning metadata. An option to copy files instead of inputting would be advantageous.

Immediate priorities

Enable the upload of data and incorporate analysis tools. The immediate priority is to have data uploaded and in use.

Security is also a priority: as soon as data are being uploaded there needs to be reassurance/demonstration that adequate security arrangements are in place.

Scientists will want first to upload their own data and programmes to see how they work, and to ensure the processes are acceptable, before using others' data. This should lead to CARMEN functioning as a working scientific environment.

In terms of data priorities, CARMEN needs to ensure that it offers a safe, managed, and well-organised environment to which one can return in 3 to 5 years in the knowledge that the data deposited have integrity and are still usable. No other data curation issues spring to mind.

In terms of where it needs to go, the project is losing momentum and needs to open up. An early opportunity to upload data and 'play' with the system is overdue. It would provide a means to find out what works and what doesn't, which would in turn lead to the identification of user priorities for work that is outstanding. At the moment the overall product still feels rather nebulous.

Sustainability

Sustainability will depend on CARMEN having been well tested and proven, so it needs to be in active use very soon.

The option of CARMEN becoming a global service would bring significant benefit, but this could prove difficult to manage in terms of security, access control, etc. It would require the development of other nodes, too, which would create new costs.

CARMEN should aim to be a global centre. Neurophysiology is a global community and should be able to sustain further CARMEN nodes in other countries. This would, however, lead to greater management overheads.

The option of CARMEN becoming a global service does increase the risk of data predation by more anonymous members of the community, yet meaningful use of the data really depends on active collaboration to enable valid interpretation and the creation of high quality analyses. This is the safeguard. The long term value of CARMEN will depend on the creation of a sustained community of collaborators. The fact that in the USA the NIH has mandated the sharing of data has been taken as a signal that the culture is changing.

What has CARMEN taught about data/data curation?

That it's a complex matter. It has reinforced the importance of data curation to science. There is a sense of relief that CARMEN will make it easier to deal with an overwhelming volume of data being generated and analysed.

It has caused researchers to think about digital curation issues such as migration across versions; the need to sustain raw data for re-use; the organisational benefits from having a standardised approach to data recording and description (better than a lab book) that will immediately show what data are missing and required.

Prior to working with CARMEN, ontologies and the ways of looking at different forms of data had never been a consideration. CARMEN has also opened new horizons for collaborating in order to 'get behind' the data.

Part 5 – Conclusions and acknowledgements

14. Conclusions

Project conduct and management

New technology solutions within the HEI environment, where the level of IT literacy remains mixed and there is a residual and critical distrust of change, are more likely to succeed if they reproduce familiar practices and tools.

All members of the CARMEN consortium are involved in knowledge production in one way or another and can demonstrate authority in their particular areas of expertise. This encouraged the development of a positive project dynamic, through the inculcation of trust and parity of esteem across the participating disciplines. However, whilst there were no *spectators*, the level of engagement has varied, with some project members holding to a limited agenda. This lack of *buy-in* has been partially responsible for delays to the launch of a functioning service, as well as for nurturing the introduction of biases by the enthusiasts and early adopters.

Much of CARMEN's strength has been due to its information specialists' willingness actively to engage with the neuroscience members of the project. Whilst necessarily providing leadership to the initial phases, the informaticians refrained from imposing a technology solution, which instead has been built upon an appropriate and extensive definition of user requirements.

As an eScience project, CARMEN may be judged an exemplar from its effective translation of technology opportunities into new services that meet the requirements of its community, which has already acknowledged a real step change in experimental practice.

Strategic considerations

The challenge of accommodating a wide variety of systems requirements, even within this relatively small sub-domain of neuroscience, provides a lesson for those engaged in designing national strategies for data preservation and sharing, where any infrastructure will need to be flexible to accommodate a diverse range of specialised needs.

National strategies for data management are likely to prove of limited value without substantive restructuring in the provision of funding and other support for informatics-dependent initiatives, particularly with respect to enabling the transition from project to service status.

Metadata

There is a general consensus amongst experimental neurophysiologists, which one may extrapolate to comparable research domains, that without the provision of effective metadata to describe the experimental conditions, the data produced and retained are virtually useless.

Since science is always provisional, it is crucial to adopt routines that preserve the integrity of the data whilst facilitating the evolution of the metadata; thereby enabling the operation of new tools and techniques upon the original data, whilst also sustaining informed access and interpretation.

The construction of technology-independent metadata schemas that aim to reflect discipline knowledge and essential experimental processes, will attract support by appealing to the communities they serve as relevant and pragmatic frameworks for data and service management.

Conceptually, the design of a standard, shared language is fundamental to the linking and sharing of data, but in a domain where data are significantly heterogeneous there will be resistance to the definition of a set of generic metadata that serves everyone but no-one well.

The assignment of metadata, even when only the minimum amount of information is demanded, requires discipline knowledge, and is likely to prove burdensome, at least in the initial stages. Participating researchers need to be fully briefed and prepared for this additional workload.

Data sharing

The introduction of new collaborative methods that include more open data sharing requires the full engagement of the affected community in a process of managed transition. This is particularly true amongst communities with a tradition of closed networks, where data is exchanged on the premise of personal trust. The imposition of a rate of change designed only to effect new technological, economic or political strategies could prove intolerable and lead to failure.

Increasingly, an agreement to share data is the price to be paid for research funding, yet the concept of earlier and more open data sharing remains highly contentious amongst the experimental neuroscience community. Here, resistance to data sharing will not reduce significantly until there is change to the domain's perception of its data having uniquely high value, drawn from the practice of highly-individualised research. Recognition of this situation by the funding agencies may help, especially if new incentives to share data are introduced; the success of CARMEN as an extended global service may itself, if achieved, also serve to break down barriers to sharing.

Engagement with digital curation

Awareness or engagement with data curation issues by researchers in the neuroscience domain, as represented by the CARMEN community, is limited. Their principal focus is upon opportunities for the better exploitation of data as a means of enabling the progress of their scientific endeavour and its subsequent dissemination. The digital curation *movement* must demonstrate the benefits of curation to the practice of science before serious engagement with its precepts can be expected.

Easy access to data, the formation of productive collaborations, greater and easier data storage, as well as the perennial demand for increased citations, are the key drivers for participation by researchers in a project such as CARMEN. Presented as an achievable model, the curation lifecycle may be welcomed as providing an enabling framework for meeting these requirements.

Recognition by project management of the value of digital curation principles and objectives has provided an additional layer of granularity to the CARMEN rationale. Yet, with a mission constrained to describing data and making data available without the exercise of formal data quality controls, CARMEN does not aspire or wholly qualify as a data curation project. One may infer an urgent need for agents of digital curation to *seed* other such projects from a pre-project phase.

Whilst there is a shared belief amongst the CARMEN community in the *data priority* for providing a safe, managed and well organised environment in which data will retain integrity, and where it can be accessed and re-used time and time again, were the CARMEN service framework extended to incorporate the more complete tenets of digital curation, it would be important not to treat the user base as homogeneous.

Despite the falling cost of data storage media, it is anticipated within CARMEN that little-used data may need to be transferred from an active repository to an archive on the grounds of reducing project or service costs. This practice, whilst designed to meet an operational imperative, risks the introduction of inconsistencies to the curation regime when data sets are divided. Rigorous rules for maintaining data relationships between the active repository and the archive would be essential.

Operational and service issues

The management of chains of ownership (of data) remains an unsolved challenge and requires significant attention, not least from legal experts. Further investigation and advice by the JISC and its agents would be of value to the research data community.

The option for providing generic core cloud services, which may have provided a baseline source of income for discrete projects, has become less attractive on account of the increase in available institutional computing power. This factor should be considered in any design for a distributed national infrastructure.

Sustainability for eScience initiatives is not simply a matter of ensuring continued funding. Metrics for success will include the identification of achievable parameters for the delivery of a usable system, a user base that will continue to grow, and evidence that the project has enabled 'new science'.

15. Acknowledgements

This study would not have been able to proceed without the welcome and assistance of the CARMEN consortium membership. I would like to record my particular thanks to the following individuals, who have contributed their thoughts and observations over the past twelve months.

Stuart Baker, Investigator

Frank Gibson, Research Associate

Colin Ingram, Principal Investigator

Alastair Knowles, Project Manager

Phillip Lord, Investigator

Evelyne Sernagor, Investigator

Jennifer Simonotto, Research Associate

Paul Watson, Lead Investigator

Angus Whyte, Research Associate (DCC/SCARP)

Liz Williams, Research Associate

Appendices

Appendix 1

Minimum information about a neuroscience investigation (MINI): electrophysiology[65]

1. General features
2. Study subject
3. Task
4. Stimulus
5. Behavioural event
6. Recording
7. Time series data

The glossary table provides a definition for each checklist item in the MINI guidelines. Examples are given only to facilitate interpretation and are not intended to be a comprehensive list of the technologies that can or cannot be recorded under each section heading.

Reporting requirements for electrophysiology

1. General features
    - (a) Date and time
    - (b) Responsible person or role
    - (c) Experimental context
    - (d) Electrophysiology type
2. Study subject
    - (a) Genus
    - (b) Species
    - (c) Strain
    - (d) Cell line
    - (e) Genetic characteristics
    - (f) Genetic variation
    - (g) Disease state
    - (h) Clinical information
    - (i) Sex
    - (j) Age
    - (k) Development stage
    - l) Subject label
    - m) Subject identifier
        - i. Type
        - ii. Value
    - (n) Associated subject details
    - (o) Preparation protocol
    - (p) Preparation date
3. Recording Location
    - (a) Recording Location structure
    - (b) Brain area
    - (c) Slice thickness
    - (d) Slice orientation
    - (e) Cell type
        - i. Target cell type
        - ii. Confirmed cell type
4. Task - if appropriate
    - (a) Protocol
    - (b) Sensory conditions
    - (c) Equipment

[65] Reproduced from http://precedings.nature.com/documents/1720/version/1/html

- (d) Recording

5. Stimulus - if appropriate
    - (a) Protocol
    - (b) Sensory conditions
    - (c) Solutions
    - (d) Equipment
    - (e) Recording
6. Behavioural event - if appropriate
    - (a) Event
    - (b) Equipment
    - (c) Recording
7. Recording
    - (a) Protocol
    - (b) Conditions
    - (c) Containing device
    - (d) Solutions
    - (e) Solution flow speed
    - (f) Equipment
        - i. Electrode
        - ii. Electrode configuration
        - iii. Electrode impedance
        - iv. Amplifier
        - v. Filter
        - vi. Filter settings
        - vii. Recorder
8. Time series data
    - (a) Data format
    - (b) Sampling Rate
    - (c) File location

Appendix 2

Example CARMEN training event

Workshop on Web Services and Workflows

What	Two day workshop providing consortium members with opportunities to try out and discuss web service technologies in a safe environment!
When	Jul 31, 2008 09:00 AM to Aug 01, 2008 05:00 PM
Where	Newcastle University, King George IV Building (Computer Cluster), Newcastle Upon Tyne, NE1 7RU
Contact Name	Suzanne Carlton
Contact Email	carmen-enquiries@ncl.ac.uk
Contact Phone	0191 222 5689 or 07922 110 362 from 1st July 2008
Add event to calendar	vCal iCal

Overview

This event is for CARMEN consortium members and aims to provide training and information on converting analysis code into web services. The concept of analysis workflows will also be introduced and discussed.

We encourage consortium members to attend.

Day One will begin with a tour of the planned CARMEN user interface, giving users an opportunity to comment and feed suggestions and improvements into the specification. The afternoon will consist of an introduction to web services and a chance to work with hands on examples. An evening meal will be arranged in Newcastle City Centre.

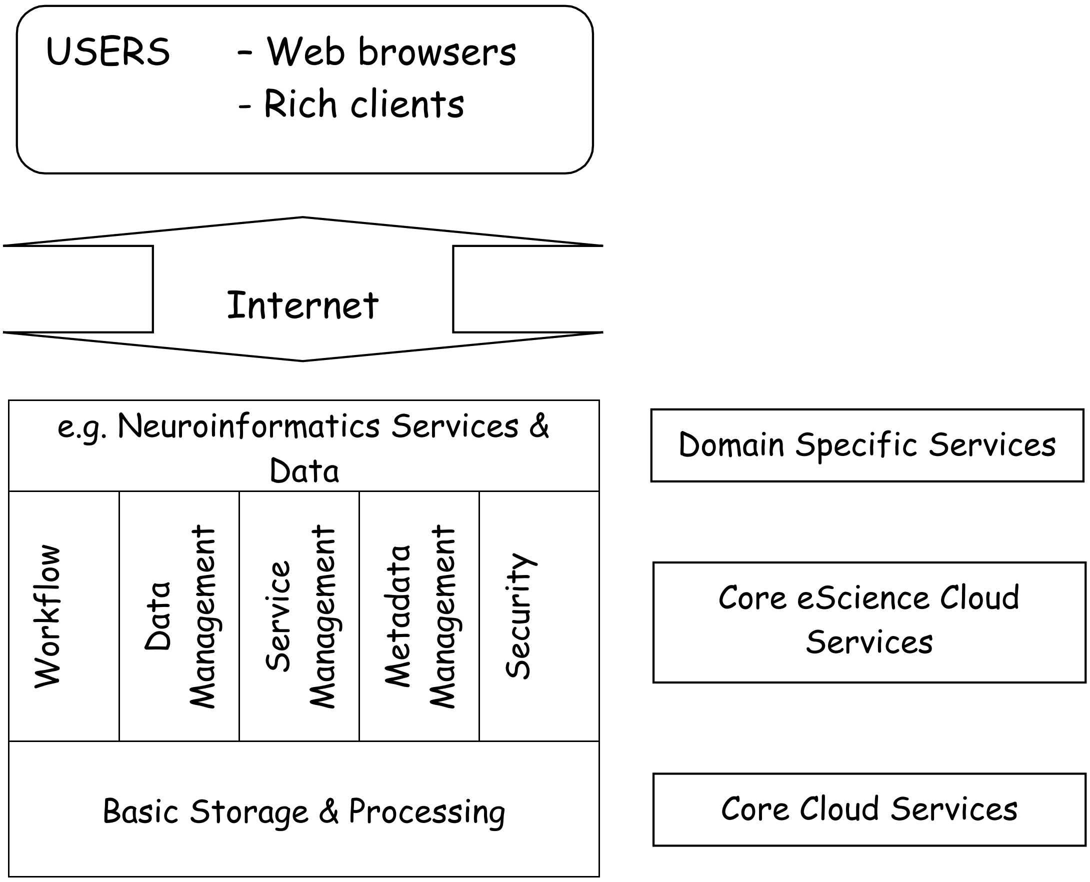
Day Two will continue the previous day's discussion on web services, and then moves onto scientific workflows, explaining what they are, what they can be used for, and why they are useful.

Accommodation

Rooms have been reserved at the New Northumbria Hotel in Jesmond for the nights of the 30th, 31st and 1st of August.

Appendix 3

The CARMEN eScience Cloud [66]



[66] Diagram adapted from Paul Watson & Jim Austin, *CARMEN: a Scalable Science Cloud*, Seattle, 14th June 2008