Effective sharing of neuroscience datasets: what are the problems?

Leslie S. Smith

Computing Science and Mathematics

University of Stirling

Stirling FK9 4LA

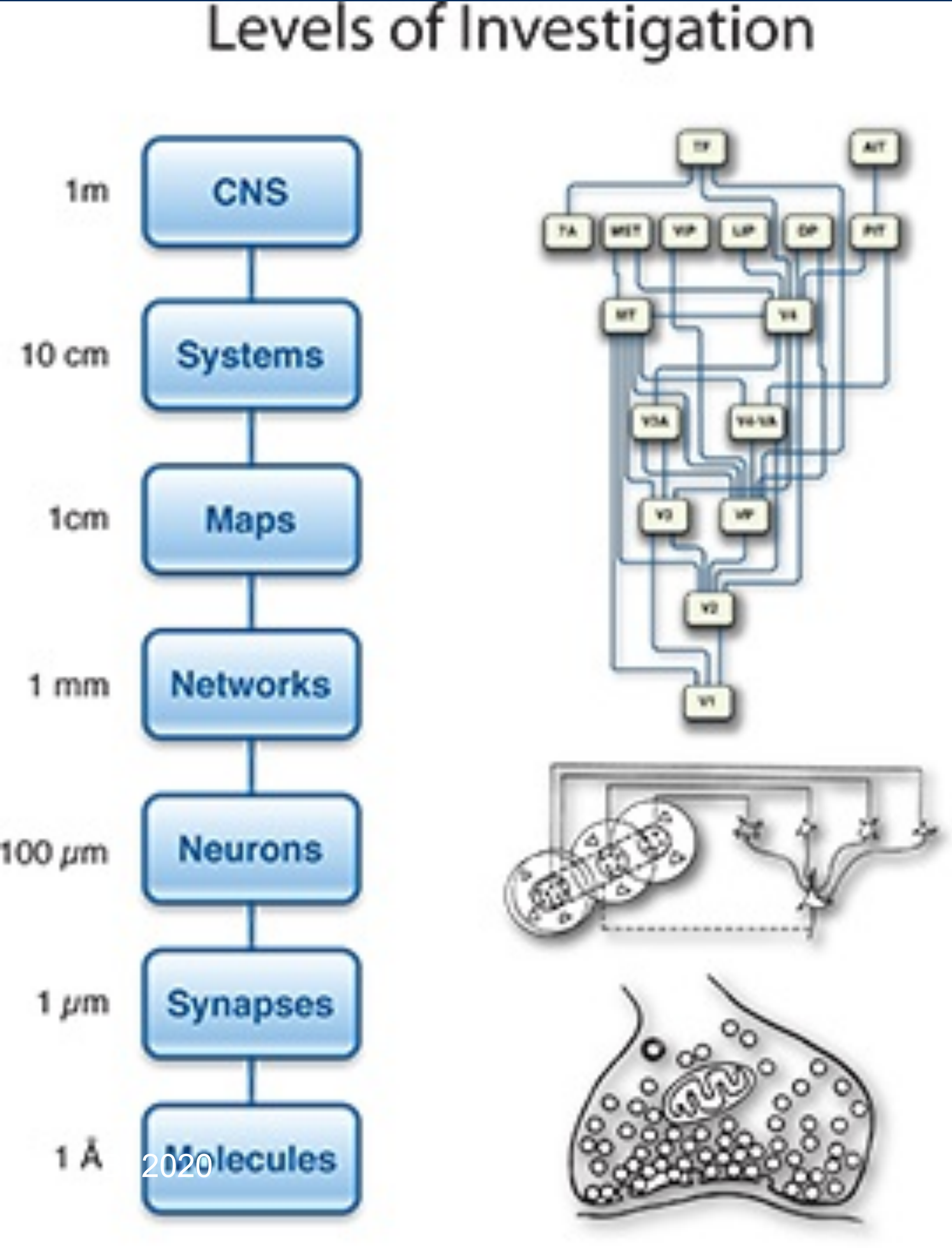
Scotland UK

Email: `l.s.smith@cs.stir.ac.uk`

UNIVERSITY OF STIRLING

Outline

- How I got involved in data sharing in Neuroscience The range of data in neuroscience
  - And why sharing is important but difficult
- The CARMEN project: sharing neuro-electrical measurements
  - What it aimed to do
  - What it managed to do
  - What was learned
- International collaboration: the INCF, NWB, NIF, HBP and eBrains….
- Where we are now: old and new problems



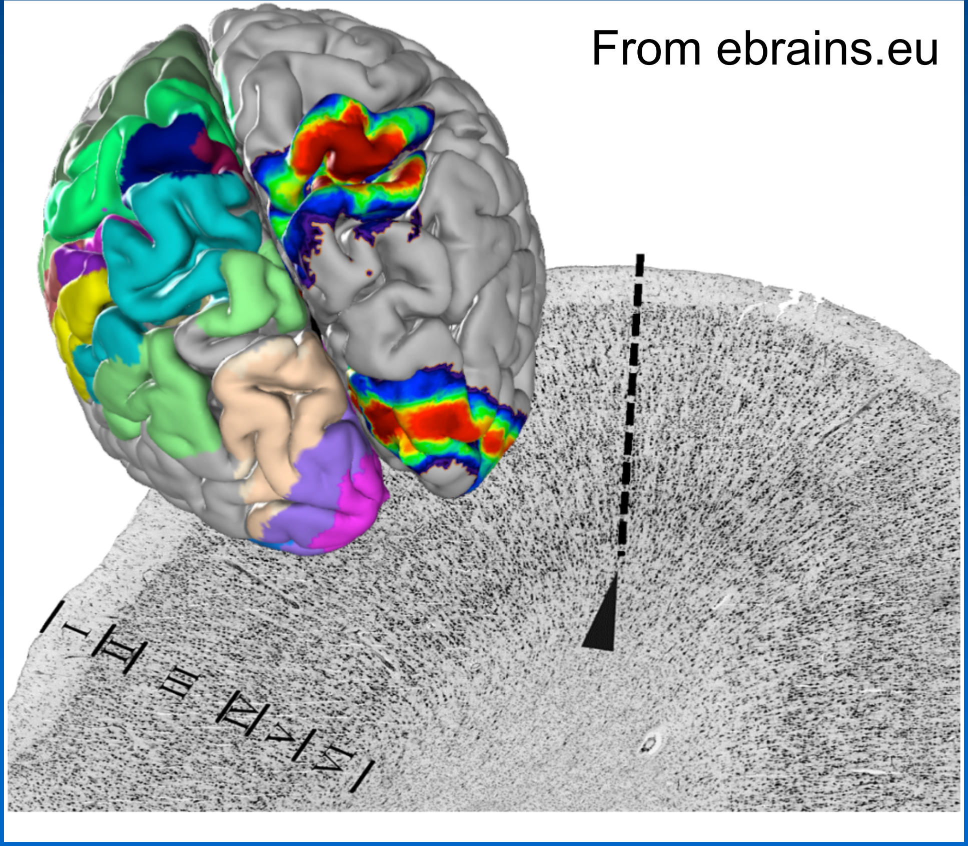
- Image from Sejnowski lab at Brown University

Why data sharing in Neuroscience matters.

- Data sharing matters for replicating results
  - Repeating experiments
  - Repeating experimental data analysis
  - Doing cross-experimental validation
- Physicists, chemists and astronomers have been sharing data for decades
- Neuroscience is hard
  - No two animals, no two brains are the same
    - Worse, brains adapt and alter continuously
  - No two experiments *ever* generate the same data!
- And that makes data sharing even *more* important!

Neuroscience data

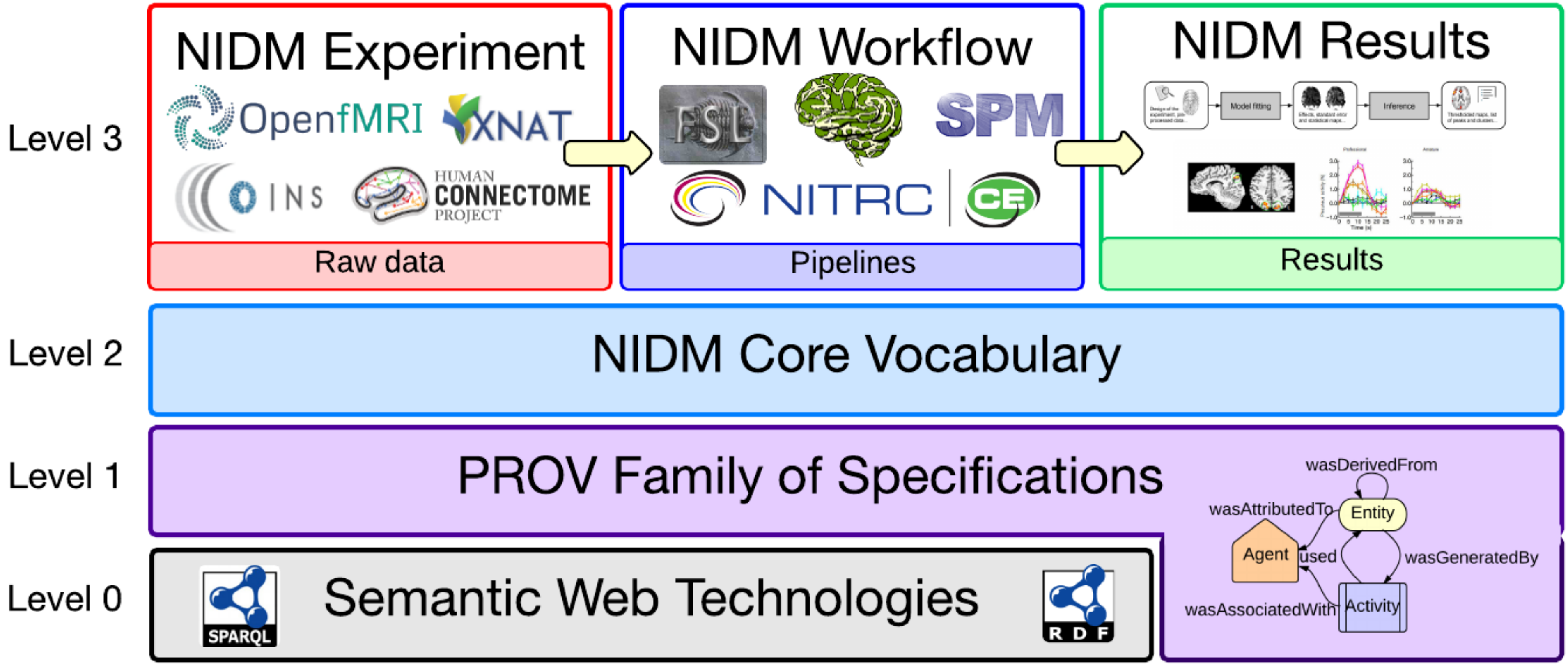
- A huge range including
  - Atlases
    - Multilevel: from global brain structure to connections between neurons
  - Neuronal morphology, microstructure, nanostructure
    - From axon and dendrite shape to vesicles and their release, to ion channel structure and operation
  - Connectomics
  - Neuropharmacology
  - Electrical/ionic movements…

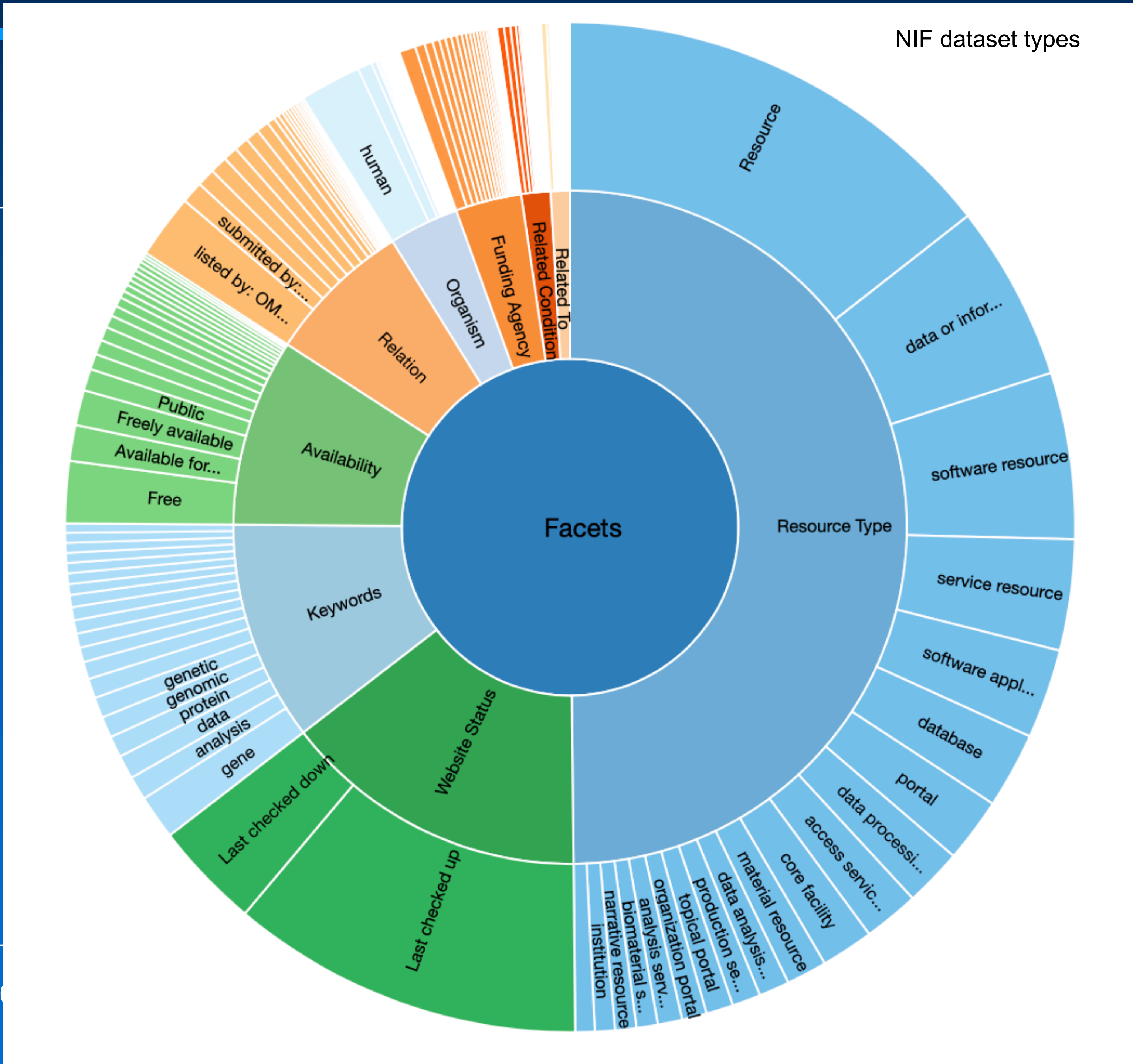


Is sharing difficult?

- Idea goes back a good way:
  - INCF (see later) established 2005
  - Dan Gardner and Gordon Shepherd's Neuroscience Database Gateway 2004
- There's a lot of databases:
  - See https://en.wikipedia.org/wiki/List_of_neuroscience_databases
  - In some areas sharing is more advanced (Neuroimaging)
  - And there is a database of databases from Neuroscience Information Framework
    - See https://neuinfo.org

Neuroimaging data model



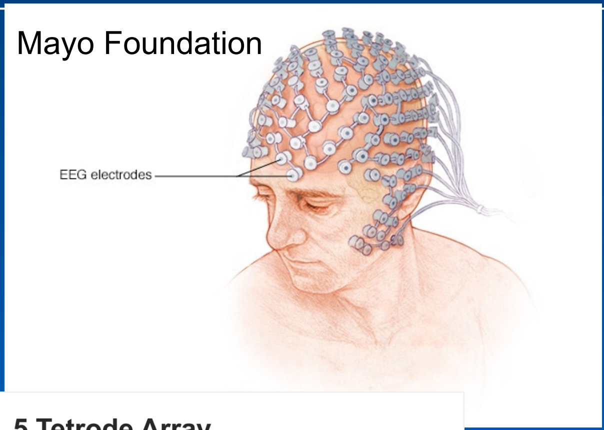

NIF dataset types
Facets
Resource Type
Resource
data or infor...
software resource
service resource
software appl...
database
portal
data processi...
access servic...
core facility
material resource
data analysis...
production se...
topical portal
organization serv...
analysis serv...
biomaterial s...
narrative resource
institution
Website Status
Last checked up
Last checked down
Keywords
gene
analysis
data
protein
genomic
genetic
Availability
Free
Available for...
Freely available
Public
Relation
listed by: OM...
submitted by:...
Organism
human
Funding Agency
Related Condition
Related To

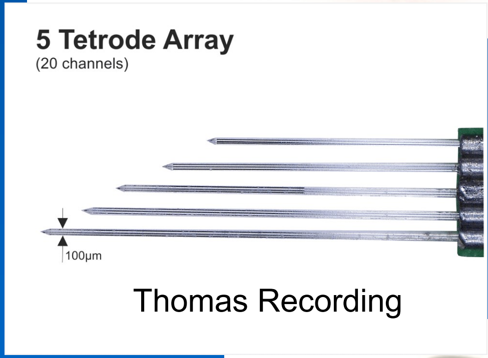
So: is there a problem?

- Vast amounts of many different types of data are generated by many labs
- Large numbers of databases for different areas
- But:
  - Sharing data introduces requirements
    - At the data format level
      - Not too hard
    - At the metadata level for the data
      - Harder
    - At the metadata level for the context of the data
      - Harder still
  - Integrating data across databases entails all the above.
- In this talk, I will concentrate on electrophysiology…

Data sources in electrical measurement

- Minimally invasive
  - EEG
    - Human, other animal
  - Tomographic: PET and MRI
    - Human, other animal
- Invasive
  - Electrophysiology (human, other animal)
    - Electrocorticography (ECoG)
    - Single/multiple electrodes; patch clamp techniques
    - Imaging-based techniques (optical electrophysiology)
- Culture-based
  - Neuronal cultures grown on electrode arrays.



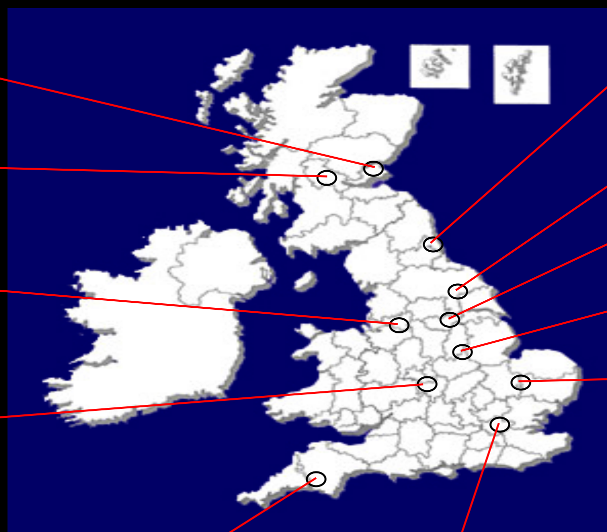


The CARMEN project: sharing neuro-electrical measurements

Code Analysis Repository and Modelling for E-Neuroscience

- CARMEN aim
  - an e-laboratory for electrophysiology.
- Project itself
  - Who we were
  - What we built and achieved
  - What we didn't achieve
  - What we learned

UK RC project: CARMEN Consortium



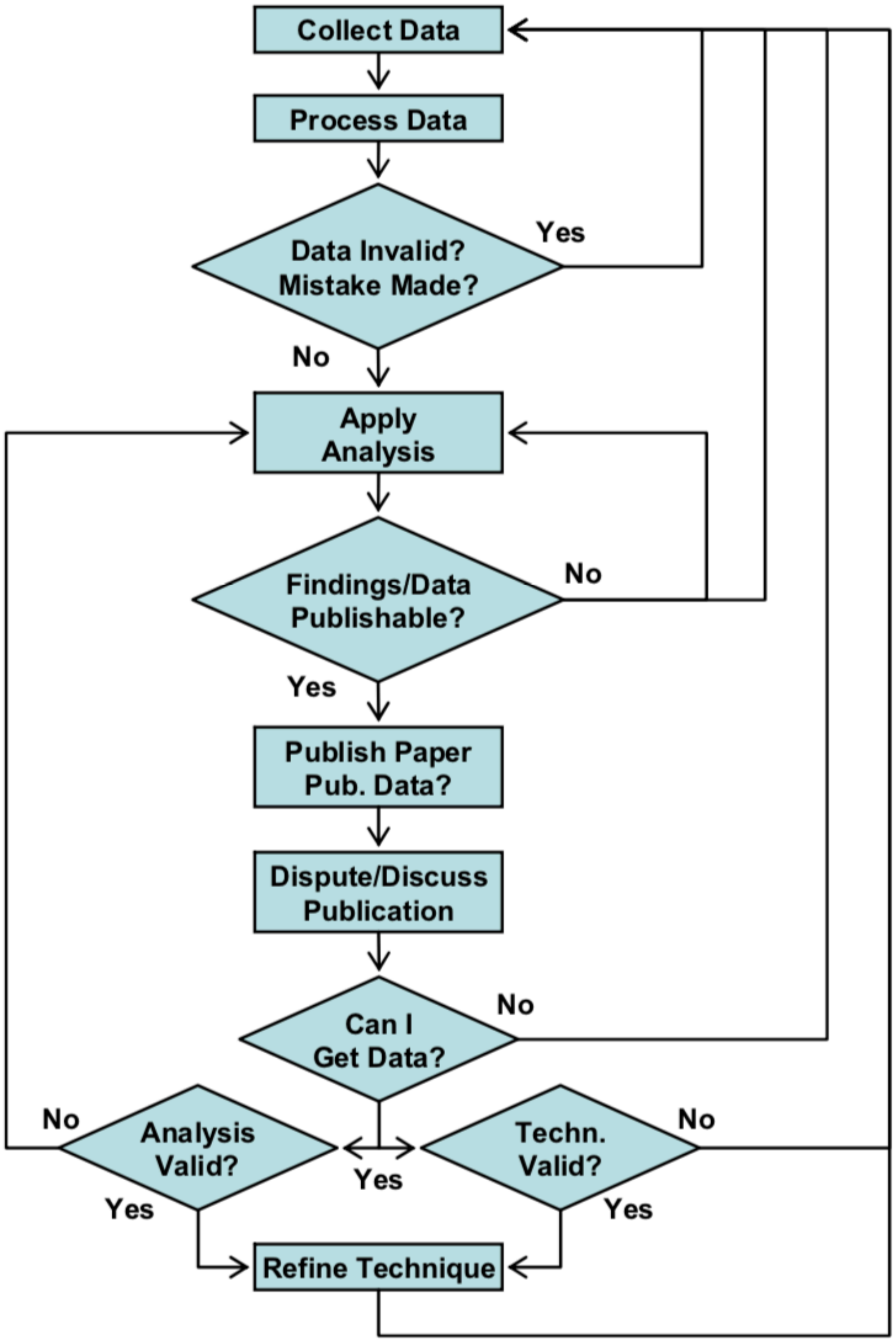
First two compute nodes (CAIRNS)

Collaborators in: Edinburgh; Berkeley; Washington; St. Louis; Aberdeen; Seoul; Pennsylvania; New York; Boston; Brazil

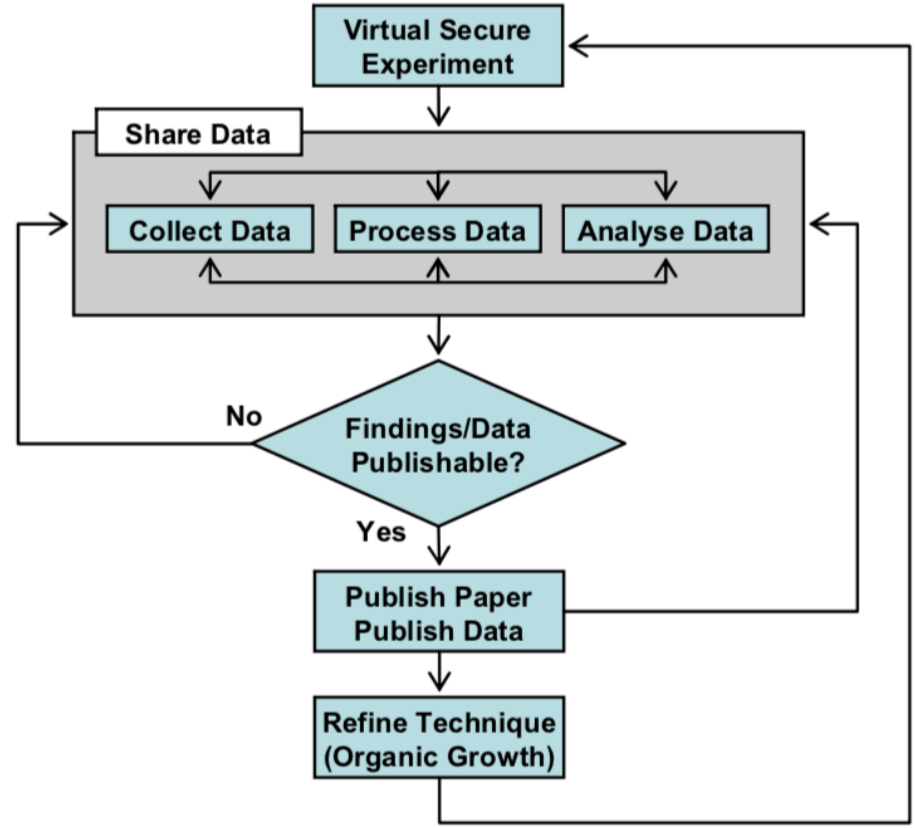
CARMEN history

- UK EPSRC funding 2006-2010
- UK BBSRC follow-on (tools and techniques) funding 2010-2015
- …but it ran out of time and money.
- The portal is no longer there
  - http://portal.carmen.org.uk
- The sudden and untimely death of Professor Colin Ingram (co-head of the Newcastle University Institute of Neuroscience) in December 2013 was a major setback.

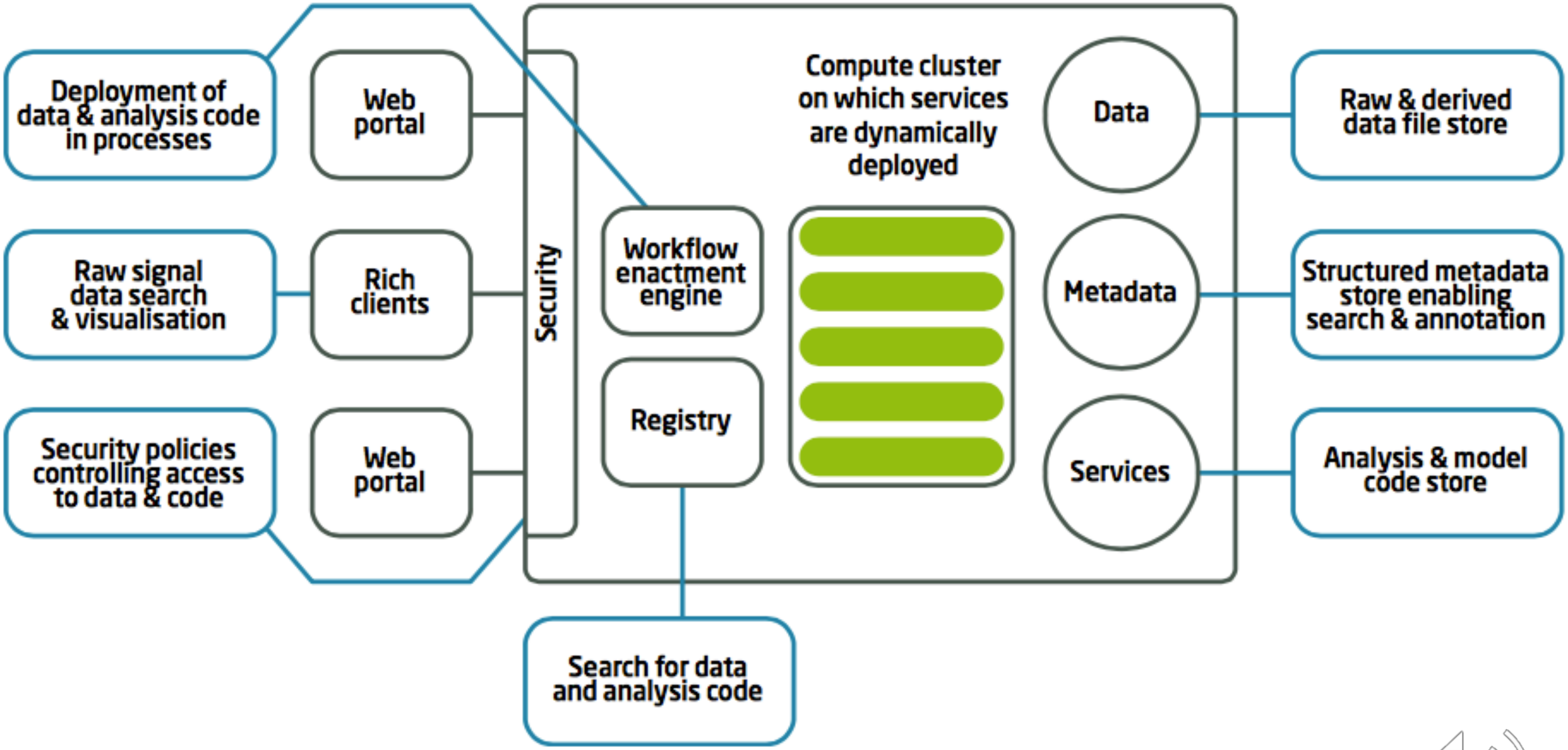
Current Neuroscience Research Process



CARMEN e-Neuroscience Research Process

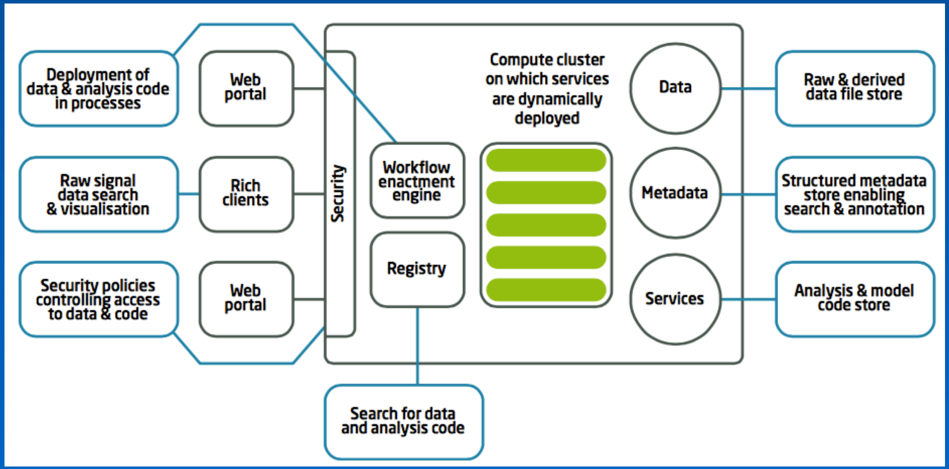


CARMEN architecture



What did we create?

- A service-based system
  - Run on machine in York University
  - Accessed through a browser
    - With additional data visualization software from York University
- Data and metadata was uploaded to the server
  - Processed there by a graphically-composed sequence of services
    - Which also updated the metadata



Issues: Cultural

- Social: not everyone wants to share their dataset
  - When should one expect to share a dataset?
    - At creation?
    - Once the researchers have published a paper?
    - Once the researchers have completely exploited the dataset?
    - Never?
  - Who should one expect to share it with?
    - No-one?
    - Immediate collaborators?
    - The research community?
- Now funders have particular policies about this!
- The CARMEN project attempted to provide suitable privacy setting to satisfy all the above answers (!)

Data (and metadata) format issues

- Raw electrophysiology data comes from many different manufacturers and sources
  - Manufacturers: Tucker-Davis, Multichannel Systems, Cambridge Neurotech, Blackrock, Plexon, …,
    - and some researchers build their own systems
  - Proprietary data formats
    - Not always open data formats!
  - Some assistance with data conversion
    - Neuroshare DLLs (unidirectional: enables data interrogation)
- Metadata needs recorded as well
  - Data source; information on electrodes, amplifiers and filtering
  - Information about the animal, about the location within the brain…
  - Contextual information about stimuli, temperature, preparation, …
- Carmen produced a document on Nature Preceding's:
  - MINI: Minimum information about a Neuroscience Investigation

Minimum Information about a Neuroscience Investigation (MINI): Electrophysiology

Frank Gibson*[1], Paul G Overton[2], Tom V Smulders[3], Simon R Schultz[4], Stephen J Eglen[5], Colin D Ingram[6], Stefano Panzeri[7], Phil Bream[4], Miles Whittington[6], Evelyne Sernagor[6], Mark Cunningham[6], Christopher Adams[6], Christoph Echtermeyer[8], Jennifer Simonotto[1], Marcus Kaiser[1], Daniel C Swan[9], Martyn Fletcher[10], Phillip Lord[1]

The problem: Data interchangability

- The CARMEN system has to handle a wide range of incoming data types as well as derived data.
  - Often unreadable unless you use vendor specific software or know the encoding format
- Data may be used by users or services or workflows.
  - In a workflow, the output of a service may be the input of the other services.
- It is impractical to have services that use arbitrary input and output data formats, particularly for workflows

–Needs data translation

–to allow resources to access a standard data format

–to facilitates an environment where data can be processed in a consistently interpretable way for both human users and machines.

Issue: Remote Data

Remote data: avoiding inappropriate or unnecessary data downloading / moving and processing:

a. A user needs to know as much as possible about data before the data is downloaded or processed.
b. A service needs to verify the data as a valid input type
c. A workflow editor needs information to pre-verify the type of the input data set from a remote data depository or output from another service in the construction of a workflow script.

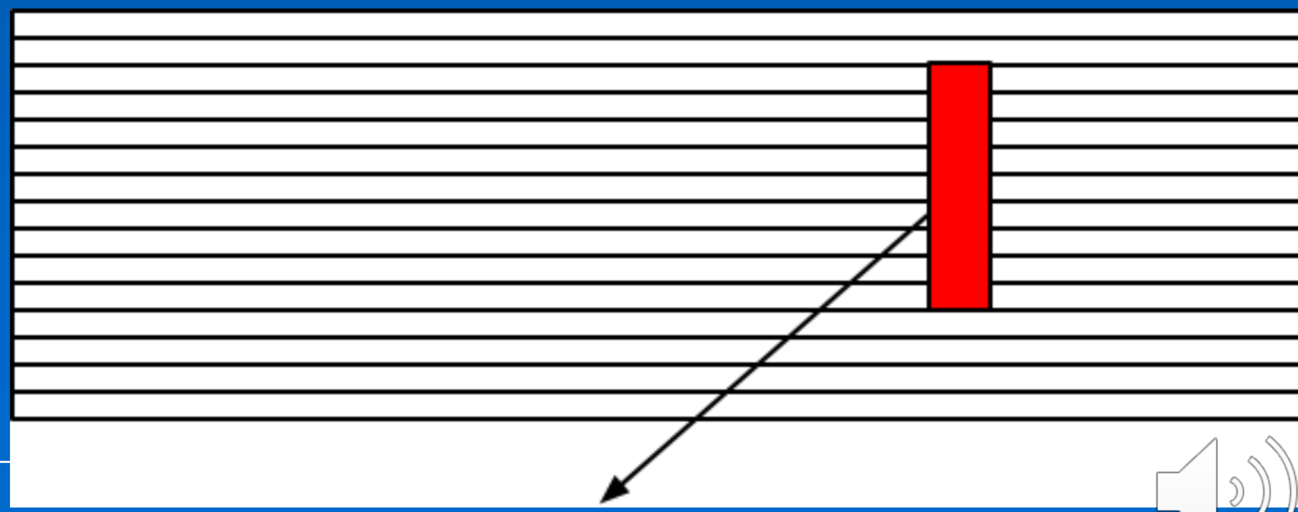
Issues:

1. How do we interrogate and understand the remote data without downloading / accessing the whole binary data set?
2. Where is a workflow editor to get information to perform the verification?

Partial data access issues

Sub-dataset selection and partial data extraction / downloading:

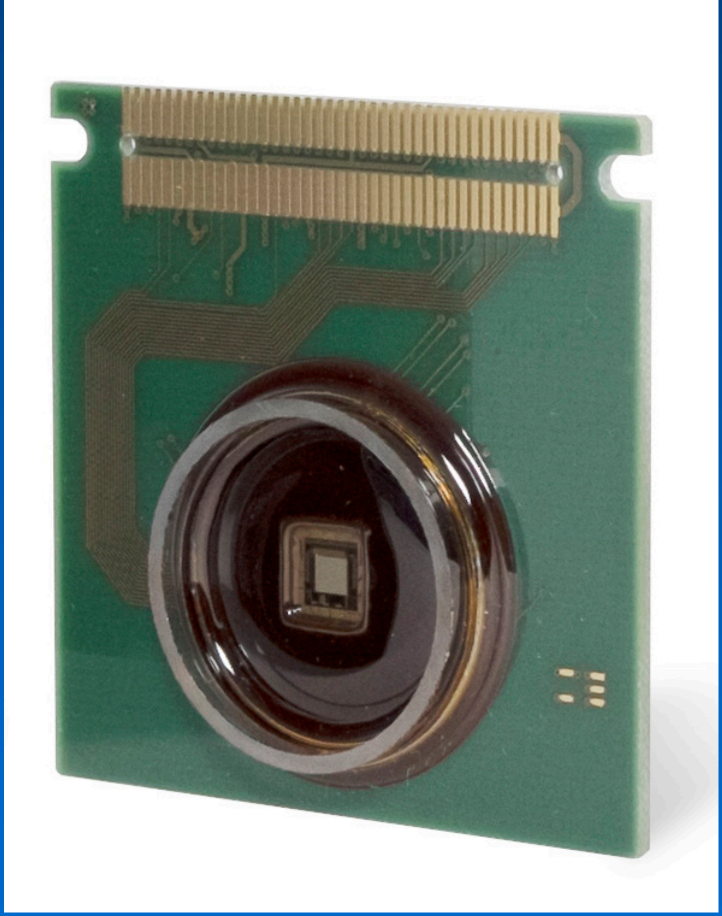
a. Neurophysiological experimental data are complex data sets. Most CARMEN services are designed to process only one of the data types within a data set.
b. Raw data contains multiple channels from the acquisition equipment but only parts of these data channels may be desired.
c. The volume of data in a channel of data may be very large but only some channels and time intervals are of interest.
d. Processed data and raw data may be mixed in the same data set.



Changing data format issues

New data types / formats are created whenever new scientific instruments or services / algorithms are introduced.

It is difficult / impossible to try to specify these precisely in advance.

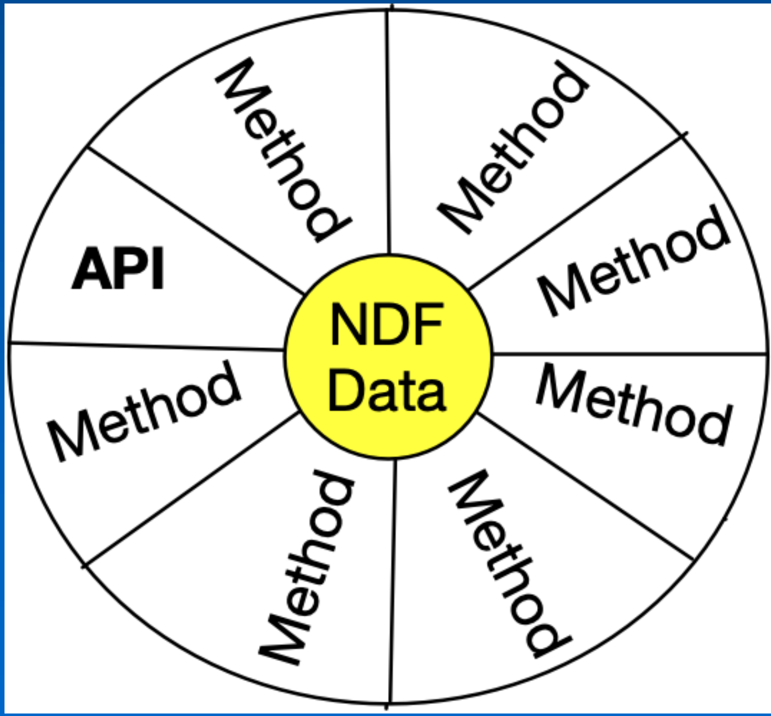


Questions:

1. Can we create services that accept new data types as input?
2. Can we create services that create a new data types as output?
3. Can all this be done in a consistent manner, using the predefined data types?
4. How can a service that uses new data types perform pre-verifying as for the predefined data types?

How to solve these problems: data/metadata API

- *Use a generic metadata system*: users do not want to use generic metadata specifications
- On uploading a data set, the metadata may not be directly available for the user – a special tool for a particular data format may be required.
- It is impractical to upload metadata manually for a huge number of data files.
- Automatically uploading metadata is equivalent to having a data standard. This implies that the metadata is already included in the data set and a data standard must be used.
- Separating the metadata from a data set affects data set portability.



Our conclusion: The metadata for the above purpose should be *integrated* with the data set: a Neural Data Format (NDF) entity (object) with an API hiding the internals.

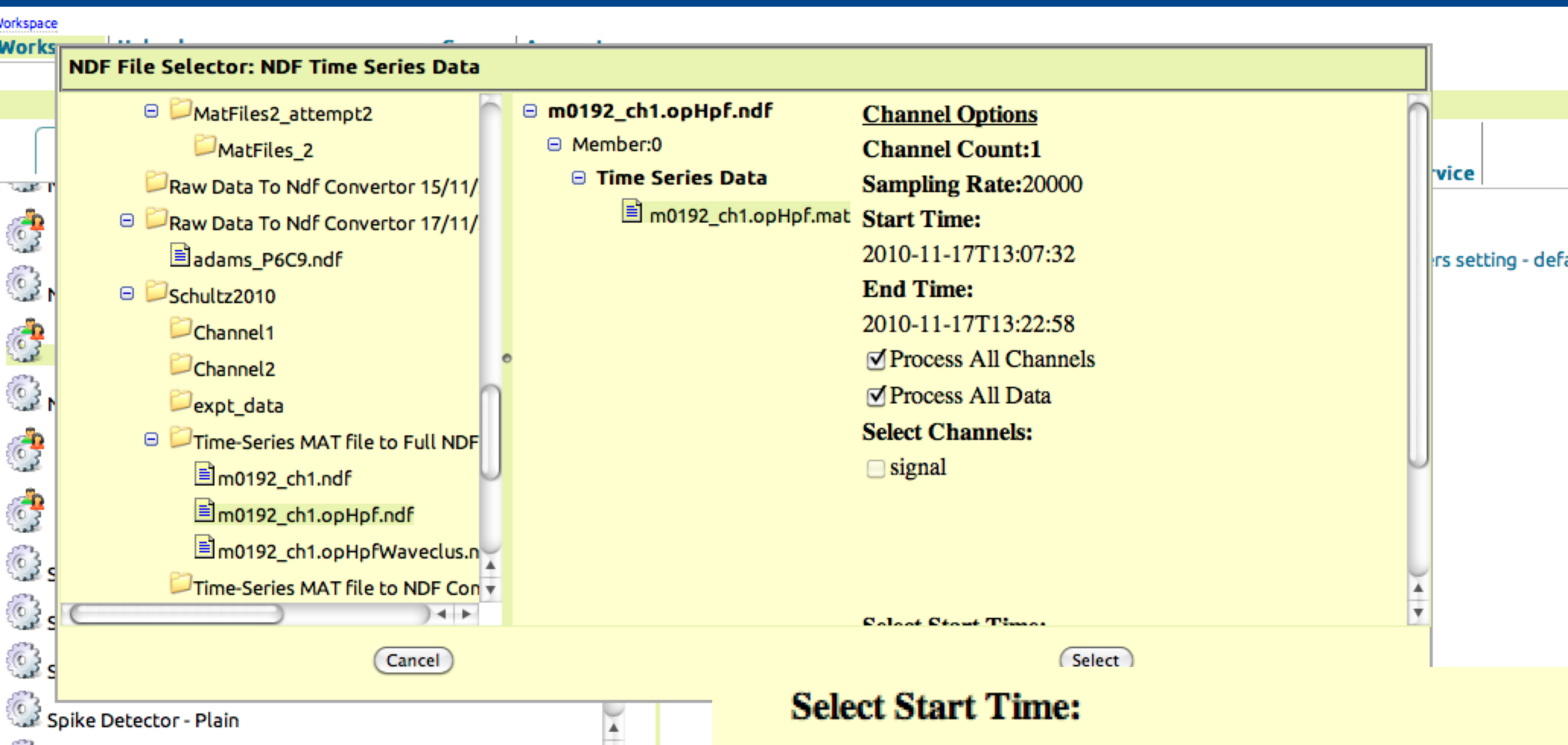
Basic data types

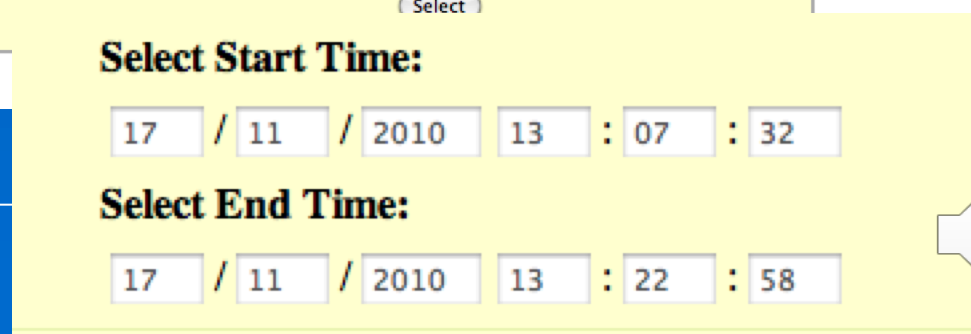
- The primary data types are
  - TIMESERIES: continuous time series.
  - NEURALEVENT: events such as spike times
  - EVENT: other event data (e.g. stimuli)
  - SEGMENT: sections of TIMESERIES data
  - GMATRIX: generic matrix data: user-defined
  - IMAGE: image data
- Since the content is described using XML, additional data types can be added to cope with new developments.
  - And the API can be backwards compatible

The NDF data format

- NDF wraps metadata and binary data together with an XML configuration file.
- Partially defined data types are extendable
- Vendor data files may also be “wrapped” as an NDF data set.
- NDF supports numerical data types from 8-bit integer to double precision floating point. This helps to reduce the data size
- NDF permits the download of data “regions of interest” (partial data access) rather than the whole data set, reducing network traffic.
- For a workflow (chain of services) a history of each process is included in the output data. This enables repeatability.
- NDF supports image data and image sequence data.
- An XML file can be used to store experimental event data, annotation etc.
- A MAT file is used as the main numerical data file format.
  - This is a publicly described data format
- Supports multiple data files for each data channel

The CARMEN Portal NDF Data Channel & Time Selector





The NDF API

The NDF API was written in C (has to be efficient):

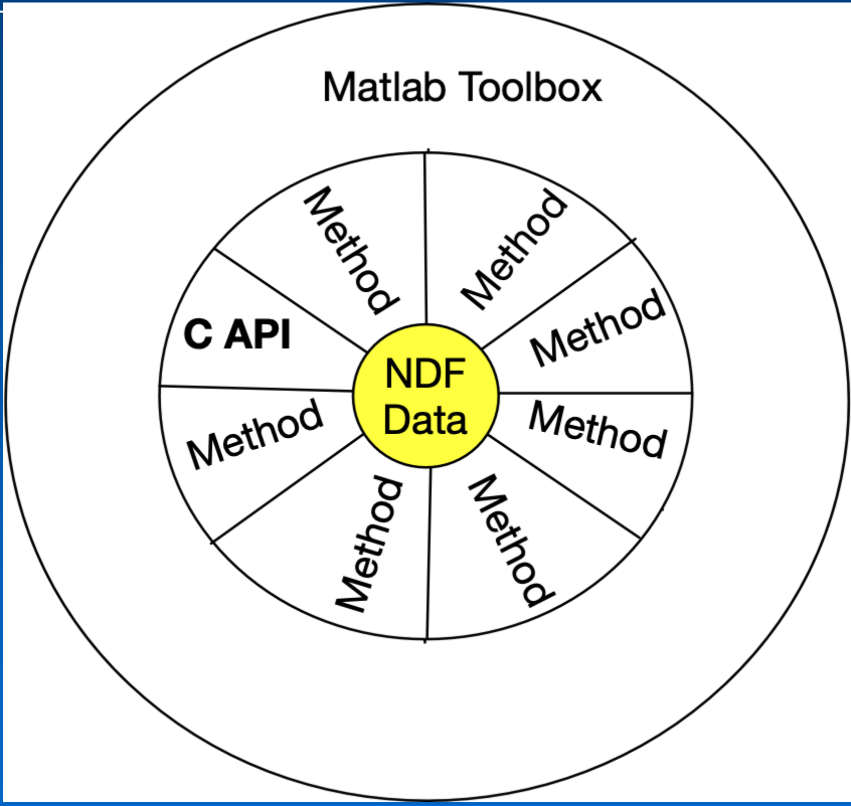
- Provides a low level I/O interface for accessing all the NDF data Translates the XML tree/node to C style data structures.
- Insulates the MAT data format and (and image format data) from the clients (so is extensible)
- Supports multiple-run data writing modes for large data sets with known total data length.
- Supports multiple-run data writing modes for data stream with unknown total data length.
- Supports zipped data stream for MAT file.
- Supports partial data reading on both compressed and uncompressed data in MAT file.
- Automatically manages the data file splitting for large data set.

The NDF MatLab Toolbox

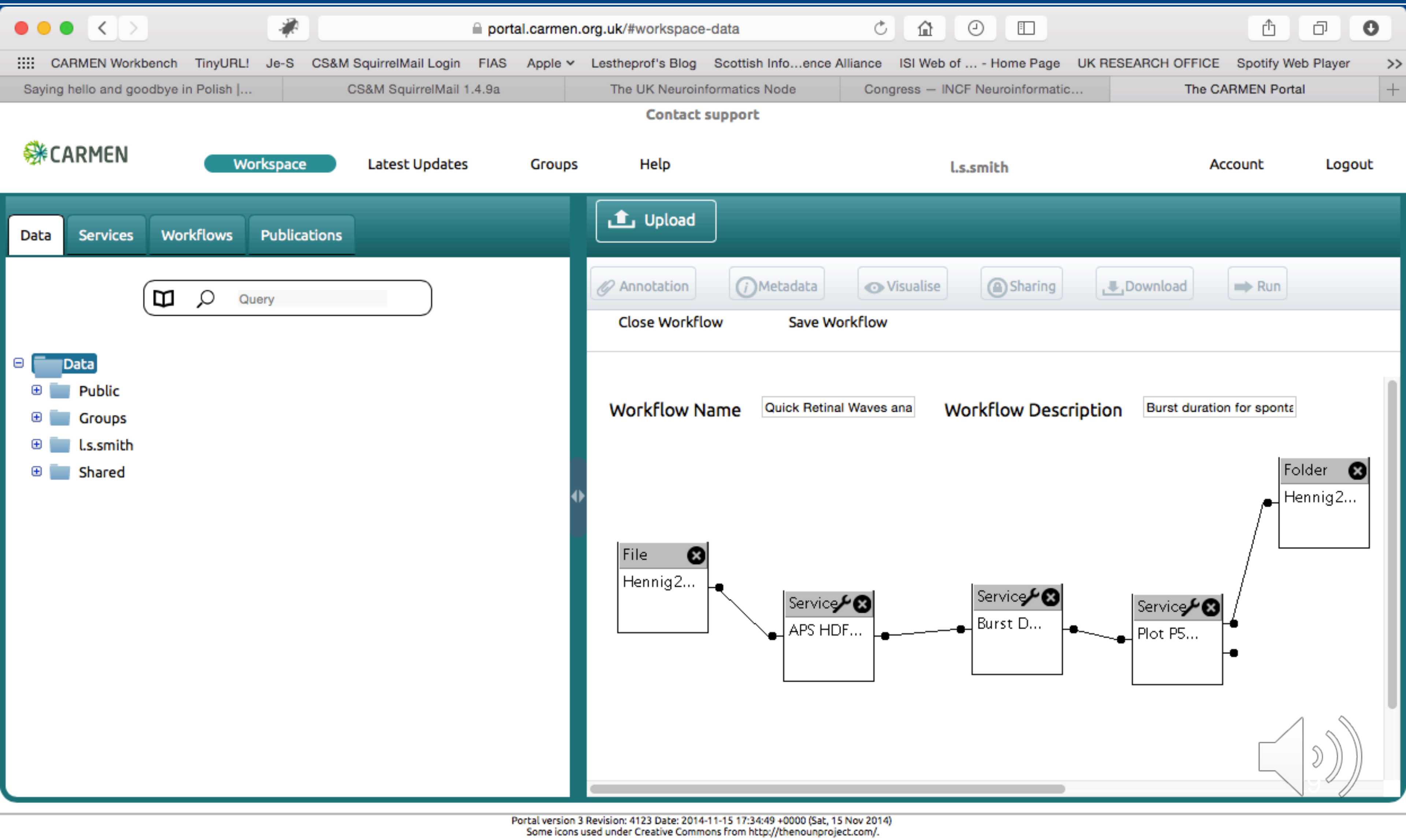
The NDF MatLab Toolbox was implemented on top of the NDF C API.

- A set of object oriented MatLab classes and functions that provide high level support for NDF data I/O.
- A “multiple data formats” to NDF converter is embedded to the toolbox as data input module.
- Full protection and auto-correction for misused data types on parameter structure.
- It has been used within the CARMEN service code programming.
- It is also used as a set of convenient tools on a researcher’s desktop for NDF data I/O and data conversion.

NDF was implemented by Bojian Liang



Workflows: joining services together



Who used CARMEN?

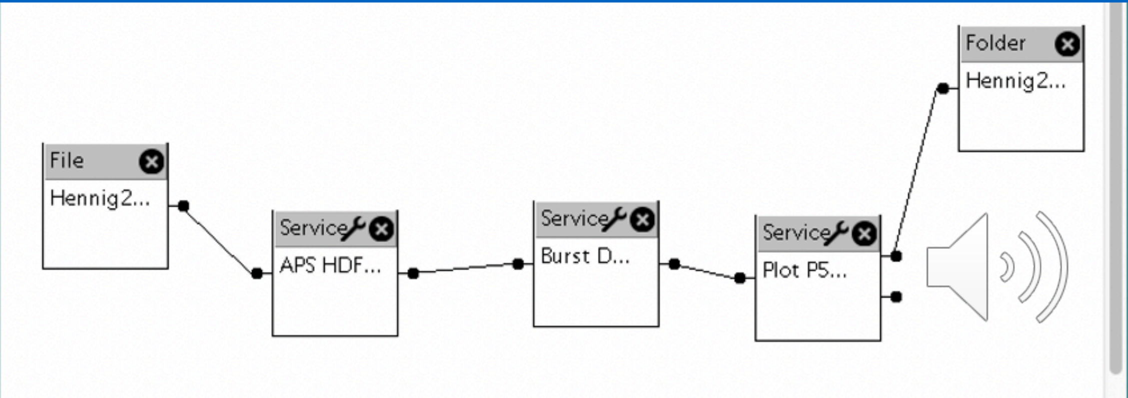
- Primarily it was used by internationally distributed research groups
  - As a way of sharing data, and of sharing some processing tools
- Some data was made publicly available
  - But not nearly as much as we had hoped

What were the technical problems?

- Large volume data movement (multiple Gbytes) was difficult and slow
  - Quite a lot of existing software did not work well under stress!
- Supporting multiple browsers proved very time-consuming
  - Providing an effective user interface meant using technologies that weren't up to scratch
    - JavaScript in 2006-10 was not well standardised
- We used a lot of Java at the client interface
  - Which then became unpopular for security reasons
- The servers we were using began to show their age
  - We didn't have money for equipment replacement, and the technology was changing rapidly.

… and other software problems

- Security was good…
  - But difficult to interpret
    - And not proven good enough for neurologists to use it for patient datasets
- Supporting multiple services was difficult
  - Written in multiple languages
    - Proved difficult to keep up-to-date
    - Not multi-threaded, not as modern in concept
    - Often based on research software
      - Not robust enough
- Creating a good user interface for services was difficult
  - Users wanted something easy to use, powerful and instant
    - We couldn’t provide all three with the resources we had.



Ways forward:
Client pull or technology push?

Client pull: the users…

- Who are the target users?
  - Clinical neurologists and neuroscientists
    - Epilepsy, traumatic injury, Parkinsonism, …
  - Neuropharmacologists
    - Assessing effectiveness of neuroactive pharmaceuticals
  - Research neuroscientists
    - In Universities and hospitals etc.
  - Neuromodellers
    - Data to constrain and test models
  - Educators
    - Training the next generation of neuroscientists

Making the system user-focussed

- What do prospective users want ?
- What do they need ?
  - What is the problem the system is trying to solve?
- What will they actually use
  - As opposed to what they say they might use?
- How can the system be made attractive and straightforward enough for neuroscientists to use?
  - What are the issues that discourage users?

Technology push

- What technologies might be helpful?
  - Note that neuroscientists don't want bleeding edge technology in their support systems
    - As opposed to their scientific systems!
- Handling large datasets: ever larger datasets!
- Remotely visualizing large datasets
- Parallelism
  - At the user level (multiple simultaneous users)
  - At the processing level (e.g. multiple datasets, or parameter searching): effective multithreading
- Search technologies
  - Searching metadata, services, workflows.

Organisations supporting sharing

- International
  - INCF
- US based (but international intention)
  - NIF
  - NWB
- EU based (but international intention)
  - eBrains
- Other interesting projects
  - Open Neuroscience
  - CRCNS - Collaborative Research in Computational Neuroscience
  - Open ePhys
- … and there are many others, particularly in MRI imaging and EEG data.

International Neuroinformatics Co-ordinating Facility (INCF)

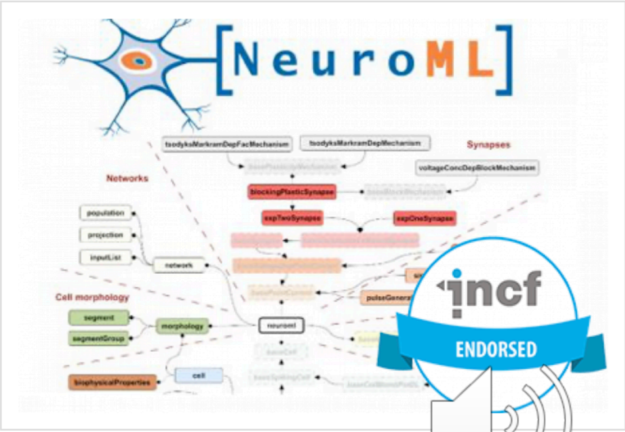
The mission of INCF is to develop, evaluate, and endorse standards and best practices that embrace the principles of Open, FAIR *(Findable, Accessible, Interoperable and Reusable),* and Citable neuroscience.

- Three countries contribute financially (Canada, Norway Sweden)
- 15 countries are associates (Australia, Belgium, Czech Republic, Finland, France, Germany, India, Italy, Japan, Korea, Indonesia, Poland, Netherlands, UK, USA)
  - Disappointingly few.

incf enabling open and FAIR neuroscience

INCF areas

- Endorsing community standards and best practices in support of the FAIR principle
  - Currently NWB (see later), BIDS (Brain Imaging Data Structure) and NeuroML (standardized model description language for computational neuroscience)



INCF Activities

- Tools and infrastructure portfolio
- Training in Neuroinformatics
- Biennial conference
- Special Interest Groups (currently)
  - FAIR Metadata Working Group
  - Neuroinformatics for cell types
  - Reproducibility and Best Practices in Human Brain Imaging
  - Neuroimaging Quality Control (niQC)
  - Neuroinformatics for Aging
  - Neuroshapes: Open SHACL schemas for FAIR neuroscience data
  - Standardised Representations of Network Structures

Neuroscience Information Framework (NIF)

- US NIH organisation
- Cataloging and surveying the neuroscience resource landscape since 2006
  - Originally led by Dan Gardner, now Maryann Martone
- Includes
  - Discovery Portal: sophisticated search capability
  - The NIF Registry: catalog of electronic resources
  - Date Sharing service: searchable collection of neuroscience data, catalog of biomedical resources, and ontology for neuroscience on the web
  - LinkOut Broker: links between PubMed articles and your data
  - Ontology Engineering: building and enhancing the main terminologies and ontologies

Vocabularies/Ontologies

NIF

- NIF has developed a comprehensive vocabulary for annotating and searching neuroscience resources
  - Critical for inter-lab co-operation
  - “a consistent, flexible terminology that can be used to describe and retrieve neuroscience-relevant resources”
- Vocabularies (and ontologies) are important in organising and then finding the correct data
  - And the greater the volume and complexity of the data the more important is its organisation.
  - See SciCrunch/NIF-Ontology, https://github.com/SciCrunch/NIF-Ontology

Neurodata without boundaries (NWB)

"Making databases about the brain more usable and accessible for neuroscientists worldwide"

- Funded by the Kavli foundation
- Neurophysiology:
  - Neurodata Without Borders: Neurophysiology (NWB:N)
  - Continuing work started in an SIG of INCF.

NWB overall aims

Experimental Design
- Stimuli
- Environment
- Trial structure
- Optogenetics

Data Acquisition
- Device settings
- Filtering parameters
- Sampling rate
- Recording/imaging area

Experimental Subjects
- Species
- Genotype
- Age
- Weight

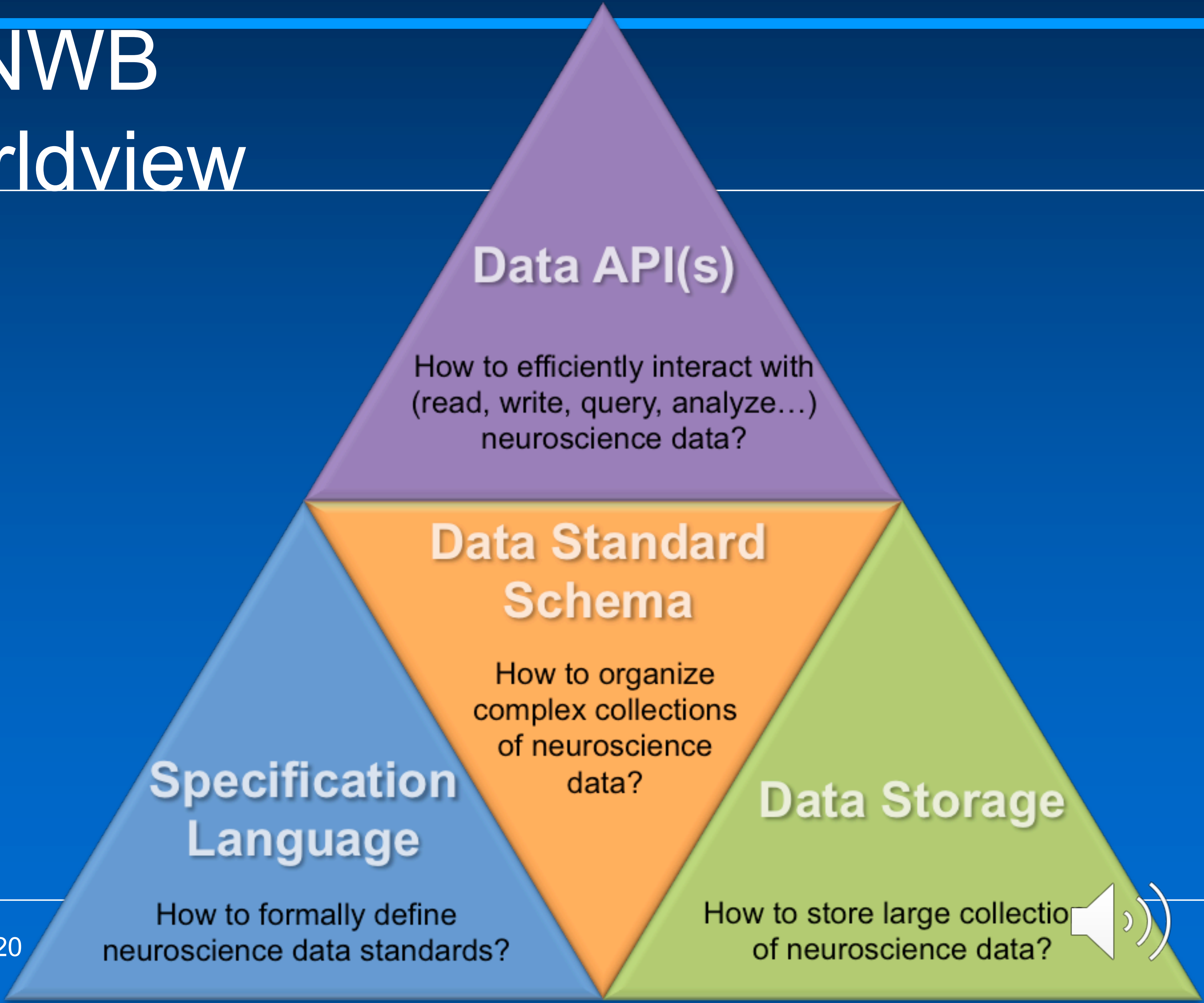
Neural Activity
- Extracellular electrophysiology
- Intracellular electrophysiology
- Two-photon imaging

NEURODATA WITHOUT BORDERS

Behavior
- Task accuracy
- Task response time
- Video and motion tracking
- Audio

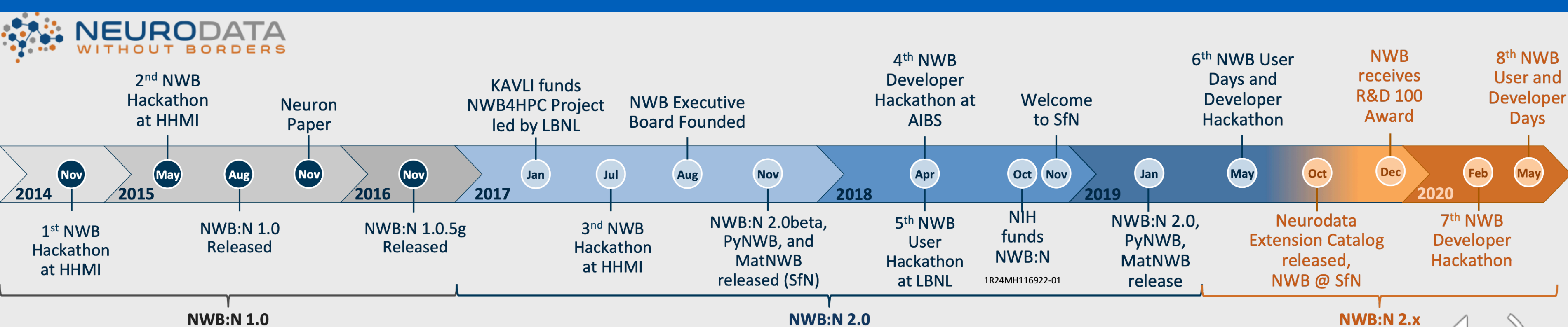
Extensions
- New data types and modalities
- Acquisition, processing, analysis, and experiment specific metadata

NWB worldview

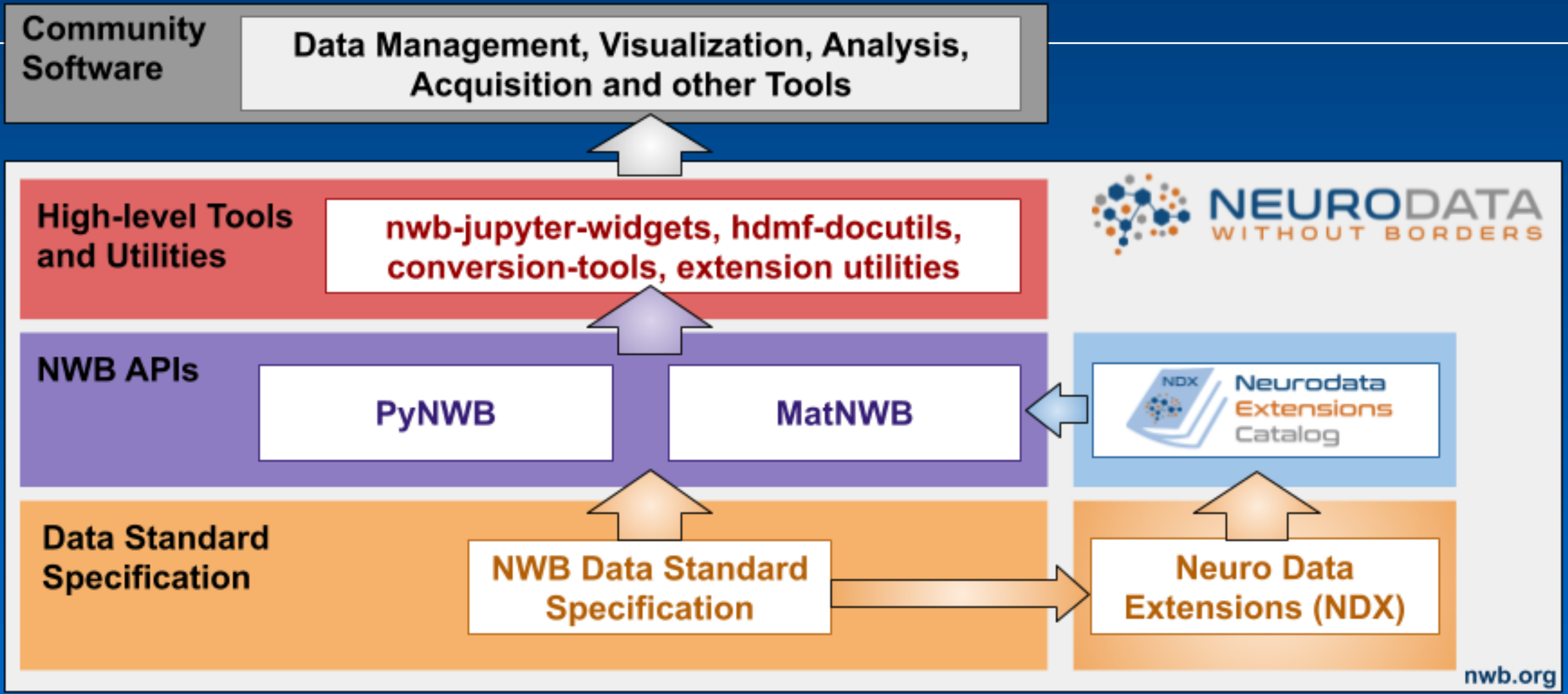


NWB:N Neurophysiology data standard

- Initial target of NWB
  - Building on from discussions that started at INCF
  - Influenced by CARMEN NDF

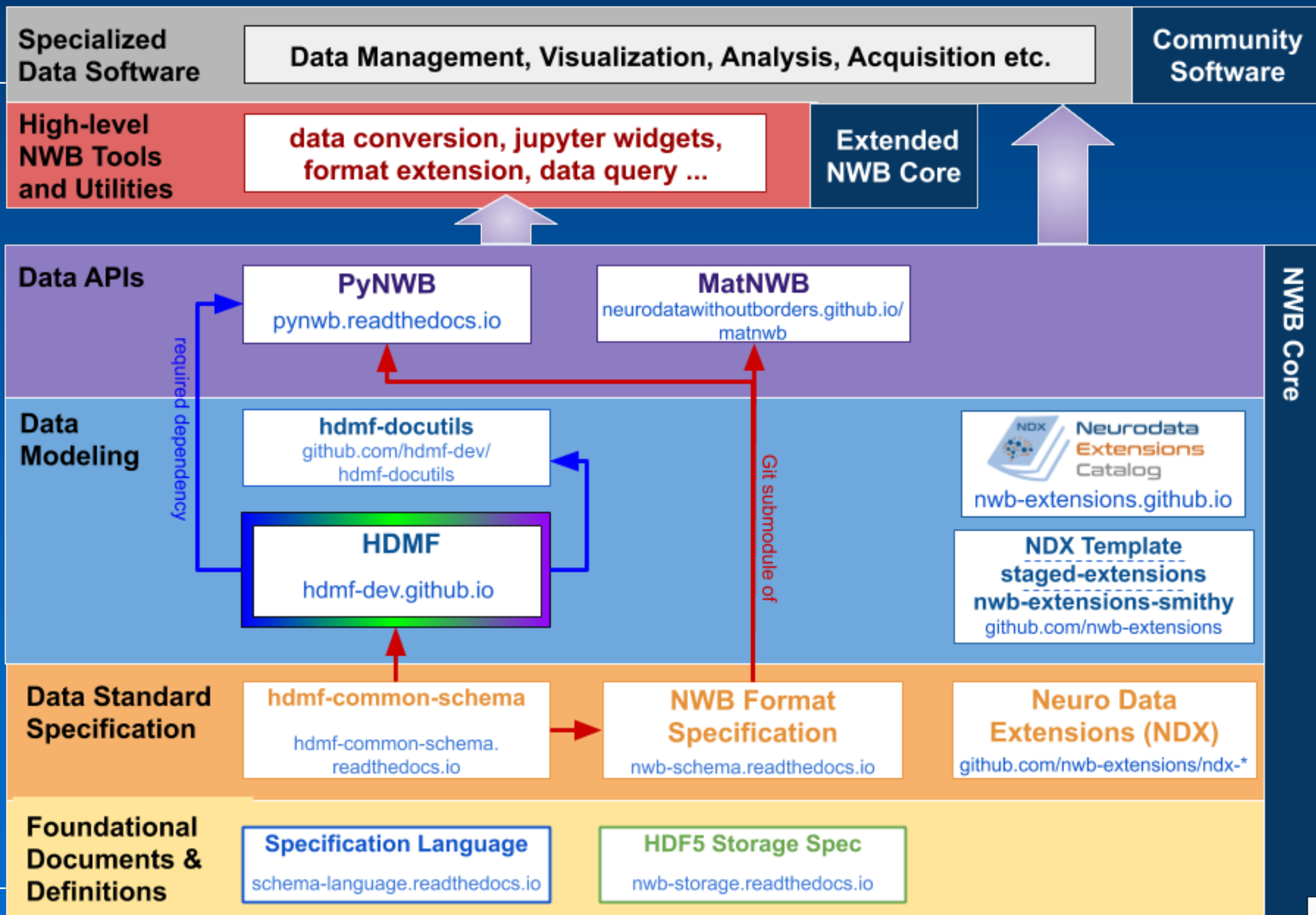


NWB:N user view



Building in data sharing from the bottom up: e.g. shareable lab notebooks, so that data is automatically captured, making sharing it easier

NWB:N developer view



eBrains

- “EBRAINS is a platform providing tools and services which can be used to address challenges in brain research and brain-inspired technology development.”
  - Services grew out of the EU Human Brain Project.
- Data sharing:
  - Behavioural data, Computational models, Electrophysiology data, Electron microscopy data, Functional imaging data, Histology data, Omics data, Reconstructions
- Brain Atlases
- Brain simulation platforms: the virtual brain NEST
- Brain-inspired computing: neuromorphic computing.

Where we are now?

Replicability, repeatability, re-usability

- Data sharing matters for replicating results
  - Repeating experiments
  - Repeating experimental data analysis
  - Doing cross-experimental validation
- Are we further forward?
  - Yes: NWB designs are better, more detailed, and easier to use
    - Publicly available “standards”
- Communications technology has moved on
  - Faster, better tools and easier to build
- … but data volumes have also increased
  - E.g. 4096 + element electrode arrays, higher resolution 3D imaging …

Where we are now?

- Analysis tool transparency and sharing
  - Open software is becoming more prevalent
  - More use of repositories
    - open repositories help produce better quality software
    - OpenNeuro
- Model sharing
  - Similar issues to above
  - Major projects to make comparability across models possible
    - The Virtual Brain, Nest,
- Re-usability of data?
  - NWB:N should help but we're not quite there yet.
  - No two brains (neurons) are the same
    - Even the same brain (neuron) will not behave in the exactly same way on different occasions
    - Need to be able to re-use datasets from different experiments to explore what stays (much) the same.
- But: organizations bringing together scientists remain fragmented.

Acknowledgements

- All the CARMEN researchers
  - Jim Austin, Frank Gibson, Tom Jackson, Martyn Fletcher, Colin Ingram, Mark Jessop, Bojian Liang, Phillip Lord, Shahjahan Shahid, Jennifer Simonotto Paul Watson, Mike Weeks

CARMEN
ENHANCING COLLABORATION IN NEUROSCIENCE

incf
enabling open and FAIR neuroscience

- The INCF Data Sharing task force members
  - Friedrich Sommer, Thomas Wachtler, Andrew Davison, Michael Denker, Jeffrey Grethe, Sonja Grün, Kenneth Harris, Colin Ingram, Marja-Leena Linne, Bengt Ljungquist, John Miller, Roman Mouček, Hyrum Sessions, Gordon Shepherd, Jeff Teeters and Shiro Usui