Encyclopedia

1st INCF Workshop on

Development of a Community-Based Neuroscience Encyclopedia

February 13-14th 2012
UC San Diego

incf

INCF Workshop on a Community Based Neuroscience Encyclopedia

Date and Location

February 13-14th 2012, University of California, San Diego

UC San Diego

Authors

Jyl Boline, Informed Minds, INCF Project Coordinator

Linda Lanyon, INCF Program Officer

Scientific organizer

Maryann Martone, NCMIR, Univ California San Diego

Sean Hill, INCF Executive Director

Participants

CHAIR: **Maryann Martone**
NCMIR, Univ California San Diego

Anita Bandrowski
NIF, Univ California San Diego

Mihail Bota
Univ Southern California

Chinh Dang
Allen Institute

Leon French
Centre for High-Throughput Biology
Univ British Columbia

Mark Greaves
Vulcan Inc.

Michael Kellen
Sage Bionetworks

Jeffrey Grethe
CRBS, Univ California San Diego

Stephen Larson
NCMIR, Univ California San Diego

Chris Mungall
Lawrence Berkeley National Laboratory (14th only)

Mark Musen
Stanford Centre for Biomedical Informatics Research (13th only)

Martin Telefont
Blue Brain Project, EPFL (by teleconference)

Sean Hill
INCF Executive Director

Linda Lanyon
INCF Program Officer

Jyl Boline
Informed Minds, INCF Project Coordinator

Acknowledgements

UCSD for hosting the meeting.

Contents

Executive Summary

The integration of neuroscience data is an extremely difficult challenge but presents an attractive opportunity to aid progress in brain research. True understanding of the brain can only be achieved through the integration and understanding of data of different scale and type. Data integration at the semantic level is facilitated by robust ontologies that cross neuroscientific domains. To aid progress in this domain, the International Neuroinformatics Coordinating Facility (INCF) Program on the Ontologies of Neural Structures (PONS), in conjunction with the Neuroscience Informatics Framework (NIF), hosted a workshop to discuss a strategy for developing and deploying a neuroscience community based encyclopedia.

The aim for this encyclopedia is to provide a semantically organized forum that is built on the latest appropriate technologies and resources for the neuroscience community to develop and maintain ontologies with clear definitions and links to relevant literature, data, models and other resources.

The attendees of the workshop agreed:

- An encyclopedia-like semantic wiki for neuroscience could be a useful method to integrate multiscale data
- It would be both feasible and desirable to integrate some of the resources represented at the workshop into this wiki
- INCF should pursue the creation of a prototype that builds on existing wiki resources; the group made several recommendations on how to proceed

Introduction

Neuroscientists collect data over a large scale, ranging from sub-cellular to the behavior of the whole organism. Each method and scale provides a limited view of the brain. In order to truly understand the brain, there is a need to integrate this multiscale data. Integration can significantly advance the field, leading to a better understanding of the brain and treatments for neurological diseases. This is a challenging prospect because the various sub-disciplines of neuroscience use different methods, data types, and terminology, and each technique gives only a partial view of the full picture. In addition, neuroscience is a quickly evolving field, the rate at which new data is produced continues to accelerate, making it extremely difficult to stay up to date on the latest developments. The complexity of brain processing in terms of highly interconnected and interdependent networks at both the microscopic and macroscopic levels adds a further unique challenge. These issues all complicate the ability to synthesize information across fields into an accurate bigger picture.

To allow neuroscientists to share information, they need access to appropriate techniques and tools that facilitate the aggregation of data, models and literature. Supporting neuroinformatics cyber-infrastructures are key for this kind of integration and to create meaningful queries. This infrastructure must provide, in an easily understandable manner, methods to share, publish, preserve, federate, access, and preferably, analyze data. It is only possible to create such an infrastructure when it is developed based on common standards. Data integration and querying functionality can benefit greatly from well-defined terminology structured within ontological frameworks; which provide a structure for common understanding and interface between different systems.

To truly integrate information from different fields into something valuable, neuroscientists need to have access to data, but also have an understanding of the context and the meaning of these data and associated information. Without this type of integration understanding remains limited to the perspective of one domain, technique or scale. Given that each field has its own methods and language, ideally, people would be involved in sharing, defining, and discussing their own data, methods, and terminology. In contrast to presentations and publications, an online forum can be used to create an enduring, more easily re-usable, expandable, and computable resource.

The challenge in defining common terms and standards is that the state of knowledge is very fluid and the way one creates definitions should ideally take account of the latest ideas, data and models. This task can be much more tractable if carried out by an engaged broad neuroscience community motivated by a rich collaborative environment that includes a range of neuroinformatics tools with value added benefit for scientists. The idea would be to provide an encyclopedic platform through which terminologies (and ultimately ontologies) could be agreed upon within the community. In addition, data in a shared dataspace could be tagged using these standard definitions and accessed via standard application programming interfaces (APIs), from the encyclopedia. Hence, data and definitions become linked. Beyond this, links can be enabled to literature and models. This provides a powerful tool for neuroscientists to probe linked data, literature and models. Potentially the platform becomes a 'living space', somewhat akin to a social media platform, in that scientists can be notified when something new happens in their domain: a new set of experimental data is released, a new theoretical/computational model announced or a new paper is published.

The aim of this workshop was to determine whether such a community-based encyclopedia for neuroscience should be developed, and to create a strategy for developing and deploying such a resource, building upon existing tools in the community. The workshop agenda is given in the appendix.

As a forum for the definition of consensus-derived terminology, this encyclopedia could become a resource that is a central component of a broader neuroinformatics infrastructure. General requirements of a neuroscience community-based encyclopedia include:

- Acts as an interface to federated literature, data and models
- Allows community contribution
- Includes and facilitates semantic annotations
- Encourages maintenance of ontologies and vocabulary
- Allows queries (preferably via ontology web service)
- Includes a rich multimedia environment: multimedia views of data and models
- Allows assisted data mining
- Includes assisted curation
- Does not restrict the state of knowledge but gives people the tools with which to use and expand knowledge

Several tools and platforms are already in use within the community and are developing at a fast pace, for example Neurolex (http://neurolex.org) which is supported by INCF. One of the goals of this workshop was to examine these and consider whether integration of these existing resources is a viable way to create the encyclopedia. The goal is to build upon these existing resources, rather than to re-invent a completely novel platform. Given the need for the encyclopedia environment to be rich enough to encourage community contribution, contain structured (formal properties) and unstructured data (text, embedded movies and widgets that can point to data or models), and facilitate querying of data and literature, a semantic wiki seemed to be the most suitable implementation platform. Members of the workshop reviewed existing related resources, and enumerated the requirements and issues for this type of next generation semantic wiki neuroscience encylopedia. They then discussed a model of coordination, cooperation and collaboration that might support a resource such as this. Members committed to provide data and platforms to begin the process to build a pilot version of the encyclopedia.

University of California San Diego - UCSD-

Concepts

1. Wiki

A wiki is a website in which content is created collaboratively by users via a web browser using a simplified markup language or a rich-text editor. Regular, or syntactic wikis have structured text and untyped hyperlinks but a semantic wiki is one that has an underlying model of the knowledge described in its pages. Semantic wikis provide the ability to capture or identify information about the data within pages and the relationships between pages, in ways that can be queried or exported like a database. Both structured data with formal properties and unstructured data, such as movies and pointers to data or models, can be incorporated in the semantic wiki and querying of this data is possible. In the context of research, the wiki can direct users to the latest datasets and journal papers. (Adapted from Wikipedia 06/12)

2. Data Federation

Through a federated database system, multiple autonomous databases, which may be geographically decentralized, can be interconnected via a network and mapped virtually into a single federated database. There is no need for actual data integration in the constituent disparate databases as a result of data federation. Rather, integration can occur at the semantic level. Through data abstraction, federated database systems can provide a uniform user interface, enabling users to store and retrieve data in multiple noncontiguous and possibly heterogenous databases with a single query. Since various database management systems employ different query languages, federated database systems can apply wrappers to the subqueries to translate them into the appropriate query languages. In the context of an international dataspace for neuroscience, the data may be abstracted to an object layer that can then be queried via services that use search terms based on an underlying ontology. (Adapted from Wikipedia 06/12)

3. Application Programming Interface (API)

An API is a specification intended to be used as an interface by software components to communicate with each other (Wikipedia 06/12)

4. Ontology

In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. In the context of database systems, ontology can be viewed as a level of abstraction of data models, analogous to hierarchical and relational models, but intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals. (Tom Gruber, in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2009)

5. Unique Identifier

With reference to a given (possibly implicit) set of objects, a unique identifier (UID) is any identifier which is guaranteed to be unique among all identifiers used for those objects and for a specific purpose. Here, the generation strategy is names or codes allocated by choice, which are forced to be unique by keeping a central registry. (Adapted from Wikipedia 06/12)

Workshop Discussions

The first day was spent examining some existing resources that might be employed in a Neuroscience encyclopedia wiki environment.

Wiki Resources

First the existing semantic wiki resources represented by attendees were examined. These included NeuroLex (http://neurolex.org), Neurowiki (http://neurowiki.alleninstitute.org), and Channelpedia (http://channelpedia.net). In brief, NeuroLex was developed for community-based curation of Neuroscience, and has been in use for over 3 years. Part of its development was shaped by the INCF PONS program and it continues to be populated through the PONS effort. It is a target for web-based data integration and any page is indexable by web search engines. Neurowiki is a semantic MediaWiki for mapping genetic instances. It combines wiki and database technology to integrate data in an encyclopedia-like format. Channelpedia is a web-based knowledge system for models of genetically expressed ion channels. It includes an interactive wiki that encourages researchers in the field to contribute both structured and unstructured data and provides a platform for the collation of experimental information from multiple outside sources (entity concepts, experimental data, etc.). This integration has enabled staff to manually extract structured information from unstructured data.

While none of the existing wikis examined completely fulfill the requirements outlined in the introduction, each of the platforms demonstrated some of the desired capabilities of a Neuroscience encyclopedia wiki environment. Genewiki and WikiPathways were also discussed as some of the most successful wikis in neuroscience. Wiki projects like this, associated with Wikipedia, likely have the largest number of readers. Hence, a Wikipedia association would be a recommended way to proceed if a Wiki is chosen as the forum for this project.

It was agreed that user interaction is one of the most important aspects to be considered: it is essential that the chosen platform be easy to understand to encourage its use. It must be built in a manner that ensures it acts as a platform for data integration and is interoperable with other resources and tools. Its technology must be stable, scalable, and interoperable with other software (e.g. Microsoft Office). It should be open source and use open standards (e.g. SPARQL, SQL, OWL, etc.). It should facilitate the integration of information from multiple sources and types, including the ability to visualize images and have appropriate links to references. There were several key methods identified as necessary for data integration and interoperability. This includes

using ontologies and standardized names for things like metadata, species, developmental stage, brain structures, diseases, and behaviors. Using standard reference coordinate systems and standard temporal references, allows the ability to map to space and time. Using unique identifiers for entities such as patient, subject/citizen, animal, dataset, aid in better understanding and longevity of data.

Content population will require input from other resources and text mining, but eventually, contributions from members of the community should be accepted (including both structured and assisted information entry). Ideally, this would also include data sharing, with simple workflow methods to aid bringing experimental information into defined data models used by the system. For others to use this shared data, certain requirements will be in place to allow a user to find it (it must be accessible through the web, include structured or semi-structured data, and use annotations as common identifiers), access and open it, which mean the data type has been specified and it is in a usable form (preferably conforming to a defined data model). Finally, semantics and the context of the data (experimental metadata) is required to understand what the data mean. Curation will be needed to review all content and ensure shared data is structured properly and accessible. Both human and automated mechanisms can play different roles in the curation process.

The encyclopedia must be easy to query from the wiki interface and via API. One of the reasons the wiki interface was preferred is because with it, the bulk of traffic can come from search engines such as Google. Another benefit of this platform is that the wiki resources presented at this workshop already have many of the desired components of a neuroscience encylopedia and can be used to aid in its development. Importantly, in order for the encyclopedia to be successful in the long run, community visibility and contribution are essential. Simple incentives spearheaded by INCF can encourage the community to build applications on top of this infrastructure.

Potential Content Services:

Several resources that provide services applicable to a Neuroscience encyclopedia wiki environment were reviewed by representatives at the workshop. These included:

- NIF (www.neuinfo.org/): a resource registry with access to 4500 resources (2000+ databases), federated data (350 million records), and literature (22 million)
- BioPortal (www.bioportal.bioontology.org/): an open repository of biomedical ontologies that provides access via Web browsers and Web services to ontologies

Atkinson Hall Building at University of California San Diego - UCSD-

- BAMS (www.brancusi.usc.edu/bkms/): an online knowledge management system designed to handle neurobiological information at different levels of organization of the vertebrate nervous system
- The WhiteText project (www.chibi.ubc.ca/WhiteText/): a corpus of manually annotated brain regions created to facilitate text mining of neuroscience literature has already been conributed to NeuroLex, any newer content would also be useful to add to the encylopedia
- Synapse (www.sagebase.org/research/Synapse1.php): an innovation space that brings together scientific data, tools, and disease models into a Commons that enables collaborative research

These resources, in addition to others (some of which are currently collaborating with INCF, such as Braininfo http://braininfo.rprc.washington.edu) all have valuable content; it would be extremely beneficial if an encyclopedia-like wiki could access this content, dynamically if possible. In addition, some of the resources have methods or technology that would also be useful if they were to be incorporated into this type of system, either to aid in population, structure, or supporting infrastructure.

These presentations brought up key questions about the interaction between the encylopedia and contributing resources. It was unclear what policies and procedures would be used to prevent redundancies, duplication of work, and to ensure syncing of changing information between the resources. The initial steps of bringing information in from the different resources presents technical problems that may be aided by using methods such as conversion to RDF and the use of unique IDs and URLs, but new tools and policies may be required. There will also need to be decisions made about what should reside on the wiki and what should reside on databases. For instance, perhaps higher levels definitions (such as those at the class

level) will make up the entries in the encylopedia and examples (or instances of these classes) be data linked to these descriptions. Policies and procedures will also need to be implemented for how to deal with different definitions for what seems to be the same entity, or what seem to be the same entities with different names. In addition, decisions need to be made about if pages will be created for every conceivable entity (without definitions, but available for someone to fill in the content), or just as someone wants to create a new one. Fairly intense curation and moderation of the initial population of the encylopedia will be crucial to laying the initial foundation, both for the content itself and for managing interoperability between the resources.

Other questions revolved around the potential users of the encylopedia and how to build for them. There are different mindsets and levels of tolerance for accuracy depending on the background of the user (e.g. clinical, basic science, military, etc), it needs to be decided if all, or only select backgrounds will be targeted for this encylopedia. In addition, there are likely to be two types of consumers; those treating it like a textbook, and those that will use data for further analysis or simulations. Questions arose about if there should be differences in the wiki for these two user types, and even if it is possible (as well as advisable and beneficial) to have a system where users impact the layout.

The final set of questions centred around the submission of data by users. The methods provided for those who want to provide information and/or data will need to be determined and vetted for ease of use. Policies and methods must be put in place for exposing and ensuring the accessibility of data coming in, as well as for handling potential errors (especially for those that might do something with the data). Finally, collation and curation are essential, but can also be the rate limiting step. Using community curation may be useful, but requires a dedicated and knowledgeable population willing to participate (for example, this work is not likely to help someone achieve tenure). It can also be difficult to get proper and sustained funding for staff to perform curation. This issue is extremely important for a forum that might be expected to have reach and longevity.

Use Cases

Two uses cases were presented to the group where an encyclopedia format might be used to integrate multiple types of datasets.

- The Allen Brain Institute has multiple atlases, including for the mouse (development and the adult) and the human; with different ontologies. In addition to anatomic and gene expression data, they have connectivity data and are beginning to branch out into other data types. Their question is how to integrate these multiple data types so users can answer large-scale questions about the brain. Ontologies can be extremely useful for connecting multiple image sets that span different developmental stages. They would like a platform that allows the ability to mine data, build models, inform new experiments and feed data back in.
- The computational modeling community is in need of resources to build models. Ideally, they would have access to a resource that defines cell types, properties, densities, relations between cells, microcircuits and dynamic links to literature and data. Computational modelers would like to know when a new assertion relevant to their model, has been made in the literature. Having ontologies to describe experimental procedures and protocols would also be useful.

Both use cases demonstrated instances where data integration and organization of these data in an encylopedia-like wiki format could be a valuable method for dissemination and access to multiple information sources.

These presentations led to further discussion of three potential use cases that can be used to drive the development of a community encylcopedia. They are:

- **Use Case 1: To organize definitions, literature, data, and models for nervous system cell types**
- **Use Case 2: To review the current knowledge, literature, data and models about ion channels in the brain**
- **Use Case 3: To define brain areas and their relationships within and across different species**

Use Case 1: Nervous system cell types

INCF and NIF have been creating knowledge bases with the aid of community input on neuronal cell types that describe and quantify existing cell class types with their properties from information in literature (including links to references). Many of the descriptions are based on morphology, location, connection information, and other key information like neurotransmitters released, receptors, and species in which it has been found. The key properties were developed based on task force meetings of the PONS program and through other organized efforts such as the Petilla terminology. The standard set of properties is accessed via a form that is invoked automatically when a cell is classified as a neuron. The information being gathered in these efforts

is useful as reference material for comparing potentially new cell types with known types and for constructing statistical representations of cell types based on their known instances. In addition, NeuroLex includes the NIF Navigator, which links to NIF's federated data resources and databases and allows one to to easily access related information, data, and literature.

As illustrated by NeuroLex, a wiki format allows the ability to add information and link to key data of all scale. This may include sub-cellular information such as gene expression, channel distributions, currents, electrophysiology traces, and models at these levels and at the level of the cell. It may also include higher-level information such as the cell type's distribution in the brain, potential functionality, models where this cell plays a role, and related diseases. In addition, this format allows contributors to define and add supporting information to the terms they use to describe a cell type. For instance, what is considered a "fast-spiking" cell may differ depending on the area and species being examined. A contributor can define what they consider "fast-spiking" and link a supporting reference, cell trace, or analog file.

A potential user of this resource might find a neuron that has the morphology of a known cell type but that exhibits unexpected spiking characteristics. If they are able to access the known information about this cell type, they may find that others have also found this behavior, a test to determine if it is in fact a different cell type, or other methods that may help them deteremine if it truly is a new cell type.

The ability to combine information across different experiment-types should allow for better characterization of cells and aid in determining if a potentially new cell type is in fact, new. Congregating this information from these different experimental areas and putting them within the context of an understandable encyclopedia format can lead to better understanding of cell types, the role they play in behavior, and what their absence or presence may mean to the function of a brain area. In addition, this format should encourage scientists to weigh in on findings and ultimately lead to a larger picture of the issues that can in turn, inform multiple areas of research.

Use Case 2: Ion channels in the brain

Channelpedia is an encyclopedia-like resource that contains comprehensive information about ion channels, such as genes, expression, distribution, function, ontologies, kinetics and models. It also includes links to pertinent literature and includes the option for vistors to contribute to the site. The main purpose of this site is to bring together key information that supports the ability to create models for each ion channel. In order to create a Hodgkin Huxley (H-H) model, they use experimental data from voltage clamp

University of California San Diego - UCSD-

experiments (found in current literature) including parameter identification, digitization, H-H model fitting, and parameter readjustment.

In addition, the wiki-like interface allows researchers to discuss and improve these models.

Further development and expansion of the kind of information along with rapid updating with new findings can greatly aid in the development of these models. Moreover, linking this information to more general cell information can help bridge the gap between ion channel characterization and higher order simulations. For instance, someone interested in modeling the behavior of a particular type of cell in multiple situations may be able to use the information found in this platform to create statistical representations of the concentration of ion channels within different parts of that cell type. This, in combination with the models of those ion channels may lead to more complete predictors of the electrophysiological activity of that cell type in different conditions.

In addition, this forum may facilitate faster information exchange between modelers and neuroscientists. It could act as an important part of validation, by being used to help clarify the question being asked of a model, and to help verify the components (from experimentalists) put into the model. It can act as a platform to help test and check the cycle between experiments and models.

Combining ion channel information with other multi-scale information in the nervous system can help link experimental and computational methods, especially crucial for multi-scale modeling, as one level feeds into another. Ideally, giving us better insight into how the nervous system works on all levels.

Use Case 3: Brain areas

Structure identification and location in the brain are key parts of all areas of neuroscience. One must identify where in the brain something is located, a function

occurs, or a disease has an effect. Atlases are used by scientists in all areas of research for this purpose. Several different atlases exist in different species, and often they are best suited for different needs. For instance, a 3D MR atlas may be less distorted by the collection process, but have lower image resolution and fewer identified structures than a 2D histology atlas. However, since 2D atlases are made from slices of the brain, there can be large gaps along the third dimension, or they may have been made using multiple subjects. A 3D atlas is more likely to have a full 3D area labeled and named as a structure, while a 2D atlas may have several instances where only the middle of a structure is identified, leaving the surrounding areas unidentified. For many of these reasons, it can be difficult to construct a full 3D structure from what has been identified in 2D atlases. Furthermore, atlases are often created using different criteria, such as topology, morphometry, functionality, development, or some combination of these. This has led to a number of different parcellation schemes of the brain along with names and groupings that are not always synonymous across the different schemes.

All these reasons make it difficult to translate between the different atlases, or can make it difficult to apply the appropriate name to new data, especially in grey areas. An encylopedia-based wiki can create a platform to help organize this information and allow a user to gain a better understanding of what atlas might best fit their needs, as well as how an atlas may translate to other atlases. For instance, the NeuroLex, in collaboration with the Structural Lexicon Task Force of the INCF PONS program, has implemented a form-based model and developed a set of atlasing tools (the Scalable Brain Atlas) that allows these properties to be defined through the Wiki. The form is automatically invoked when a structure is classified as part of the nervous system. The Scalable Brain Atlas can be used to automatically link a spatial definition of a structure to the appropriate page.

There are a couple ways to address this problem, one is purely by using semantics, the other by using a combination of semantics and spatial mapping. In both cases, key information will be needed in order to map across these atlases, such as what criteria were used to define structures, descriptions, and links to references for each structure. In addition, there should be a way to review similarity or overlap of structures across parcellation schemes, along with supporting information. If each parcellation scheme has an atlas that has been registered in 3D space, and with the appropriate supporting infrastructure, they can be used to quantify overlap between structures across parcellation schemes. For instance, a user that wants to know the relationship between a location in one atlas to other atlases, could determine the reasons one atlas calls it one area and others another, and potentially visualize overlap between the atlases in flat maps, 3D surfaces, and volumes.

In addition to acting as a space for the review and discussion of parcellation schemes, links to actual data collected from the region of the brain can be included, such as gene expression information, neuronal distributions, connectivity, structural, functional, and electrophysiology traces. In addition, roles these areas may play in circuits, diseases, and functions can also be explored with this format. These all might be used to aid in statistical analyses of the area, and in the creation of new models. Ideally, these methods may also be used to examine similarities across species. Finally, this may also link to tools where experimenters can identify where they collect data, whether they are chunks of brain, slices, or point locations (e.g. electrophysiology traces), and have the option to contribute their data back to this system and register it to a reference atlas.

Recommendations

The workshop participants recommend the creation of an encyclopedia for neuroscience that is capable of integrating multi-scale data using the semantic wiki platform. The technology of the semantic wiki has now matured to the point where it was considered to be the most effective way forward because it is easily accessible, indexable by search engines, focuses on content, and can be used to take advantage of the power of social media. It was also clear that it would be both feasible and desirable to integrate some of the resources represented at the workshop into such a platform. Many of these, including INCF-supported initiatives such as Neurolex, are already much developed in this direction in terms of content and capabilities. The aim is to capitalise on these existing resources and current developments within them. The initial content of the encyclopedia should come from NeuroLex, Channelpedia, BAMS, and the Blue Brain Project. Future work needs to be done to determine the specifics of how to integrate the recommended content into a semantic wiki and what infrastructure is needed to maintain synchronization between the encylopedia and these resources.

Component Requirements

Several components were recommended for this system; however, it was understood that not all of these can or should be implemented immediately. Those discussed include:

- Basic overviews
- Ontologies
- Search (Google-like and ontology-based)
- Links to literature
- Dynamic data

- Models
- Image and spatial tools
- Curation tools
- Tool access

General User Capabilities

There are different potential users for the Neuroscience Wiki Encylopedia as envisioned by this group. There are those who will want to find information and/or data, and those who want to contribute information and/or data. In addition, the level of interaction of both types of users with this resource could vary from a basic to a very complex level in terms of search functionality and data contribution and integration. It was agreed that the initial focus should be on providing a resource for those who want to find basic information and/or data. Simple mechanisms for people to contribute information may already exist and be easily adopted, for example the input sheets offered by NeuroLex.

It is important that the system be easy to use and understand. This could be facilitated by the use of review pages and examples given within the wiki model. The system must be fast, stable, secure, scalable and extensible. Users should be able to easily query the resource and an API should be provided to allow tools to access it. At a minimum, entries should include a summary of the information with relevant references and links. However, this environment should also allow for the seemless integration of information of different types from multiple sources. For instance, an entry might include any or all of the following:

- The actual data or links to the data if it resides elsewhere
- Embedded or links to relevant multimedia, such as images and movies
- Any models related to the topic with supporting references, parameters, etc.
- Links to supporting analyses when available (e.g. Sage's Synapse)

If the content resides elsewhere, there should be methods for viewing the content at its source. Proper attribution and possibly some level of confidence or curation should also be attached to as much content as possible.

A wiki format was chosen for this forum since it can allow community contribution, which can be made simpler by the use of structured and assisted information entry, and documentation and examples. In addition to new entries, contributors should have the ability to add1: references, annotations, data, and analyses.

Infrastructure Recommendations

The group made several recommendations for the development and implementation of an infrastructure. This was envisioned as a multi-step process, with the initial incarnation being simple, but with an infrastructure that can evolve into something more complex. To test this at each phase and to reduce the liklihood of needing to rebuild, it was recommended that the encylopedia platform be released early and often. The other recommendations focused on three areas, the wiki itself, standards, and contribution of data and/or information.

Wiki recommendations

- Create a front end that accesses both federated systems biology databases and the semantic media wiki backend
- Potential wikis to build on: NeuroLex, Neurowiki, Wikidata, and RFAM
- Integrate as much as possible with Wikipedia. Wikipedia does not provide all the semantic functionality required and, thus, is not a complete solution for this encyclopedia. However a subset of information could be contributed to Wikipedia
- Focus on methods for returning at the top of Google and other search engine searches

Standards

- Build on ontologies (design in the ontological structure from the start)
- Use open standards (e.g. SPARQL, SQL, OWL, etc.)
- Use persistent IDs whenever possible

Contribution of information or data:

- Set up procedures and policies for keeping information and data current between the encylopedia and any contributing resources (as current extensions of Neurolex and Neurowiki provide)
- Set up proceedures and a policies for tracking concepts as they evolve
- Set up proceedures and a policies for proper acknowledgement of contribution for every piece of information, and as concepts evolve (as typically provided by a wiki platform)
- It isn't completely clear when data should reside in databases associated with the wiki and when it should just be pointed to; however, it seems best to point to existing data on a different resource whenever possible (facilitate dynamic

and distributed content)

- Use as needed automated tools to generate pages
- Use data models to facilitate the handling and contribution of data
- INCF has prototyped a standard Object Model for Neuroinfomatics (OMNI) to provide object encapsulation and a common object model layer. This layer will facilitate analytical processes and the operation of services across different data types and platforms and offer outputs in a standard format
- Offer workflow methods for data contribution that aid in file handling, storage, provenance, and analyses
- It needs to be deteremined how much the community can update without curation; it is not clear where the line should be drawn between engaging the community and protecting the ontology. It may be better to start with a smaller system with a clear ontology and grow from there
- Use INCF Nodes and Task Forces to spearhead population

Content

The initial scope for the encyclopedia should provide a sufficient user base, but be limited to aid proper development and feedback. The first aim should be to find a "customer" and define minimum components and requirements. Given the attendees of the workshop and resources represented, a potential first focus is "neurons and their properties" with the following content providers:

Brain areas **(Begin with those already in NeuroLex)**	• Use Pan-Mammalian Parts list created by PONS and link to UBERON • Include ability to view areas using Scalable Brain Atlas (already possible with NeuroLex)
Cell information **(Cell types, neuron descriptions, morphometry, gene expression, electrophysiology, etc)**	• NeuroLex (currently an INCF contract with Gordon Shepherd to expand) • PONS Neuron Registry • Cell Ontology • BAMS • Blue Brain Project • Allen Brain Institute
Connectivity	• BAMS • CoCoMac (via Scalable Brain Atlas)
Ontology	• CUMBO (general terms essential to representing brains across mammals) • UBERON (multi-species anatomy ontology) • Relavent NIFSTD ontologies • Cell Ontology (cell types) • Other relavent ontologies from OBO Foundry
Experimentally generated data	• Allen Brain Institute • NIF resources • Blue Brain Project
Integrate more specialized encyclopedias	• Channelpedia

Services might be added such as the ability to generate and render surfaces or classify neurons

In addition, automatic data mining can be used to help populate content (e.g. WhiteText project). Other content can be made accessible by linking to more complex searches (e.g. link to functionality in the original resources).

To encourage community contributions, it will be desirable to take advantage of crowd-sourcing and the social network component (e.g. link to Wikipedia, have some sort of forum). Also, it will be very useful to create incentives for people to build applications on top of the infrastructure (create extensible and open infrastructure and social or financial motivation).

Next Steps: Recommendations on how to Proceed

INCF should manage this process, and provide guidance and support to create a prototype version of the neuroscience encyclopedia, starting with resources volunteered by attendees of this workshop. The initial step is for Mark Greaves and Stephen Larson to develop a next generation semantic wiki that may fit the initial requirements of the group. It should then be populated with appropriate content from NeuroLex (focus on neurons). At this point, relevant and easily integratable information from Channelpedia, BAMS, and the Blue Brain Project (cell type information) should also be brought into the forum.

Once this has been created, it should be evaluated by as many of the attendees of the workshop as possible, along with a small group of neuroscientists and ontologists. At this point, recommendations for improvement and a concrete plan for development, funding, and population should be established.

Outstanding issues:

- As this encyclopedia and NeuroLex will have many overlapping features, it was left for future discussions if and how these two should coexist. Once this has been established, the procedures for this evolution will be put into place.
- The meeting did not draw a conclusion about branding the new product and naming it. INCF should consult on this topic at the prototype evaluation stage.
- An evaluation plan is needed to assess usage levels (note, using publications to assess usage is not ideal, due to the long time lag). INCF should draft a proposal for the plan and have this reviewed during the prototype evaluation stage.
- Right now, each project is performing the curation for its own tool. Should these individuals continue this curation or should they be working jointly on a shared resource? What have we learned from current curation patterns from these different groups?

As part of its PONS Program INCF will consider how to guide and move this project forward beyond this initial plan. It might be useful to create a small working group that includes representatives with the following expertise (and possibly others as the project matures):

- Semantic wiki development
- Neuroscience content providers
- Ontologists
- Neuroscience experts in the targeted content areas

Appendix

Program

DAY 1 : REVIEW OF EXISTING SERVICES & CREATION OF A SHARED VISION

Monday 13th Feb, Room 2004, Atkinson Hall Building

1pm	Arrivals and lunch (Contact Point: Linda Lanyon)
1.30-2pm	Introduction & Welcome Maryann Martone
2-2.30pm	Requirements of a Community Based Neuroscience Encyclopedia Sean Hill
2.30-3.30pm	Current Wiki Platforms (20mins each): Stephen Larson Mark Greaves Sean Hill
3.30-3.45pm	Coffee Break
3.45-5pm	Overview of other existing services (10mins each): Anita Bandrowski Mihail Bota Leon French Mark Musen Jeff Grethe Mike Kellen (tbc)
5-5.30pm	A shared vision: Summary & setting the scene for day 2 Sean Hill & Maryann Martone
6pm	Tranport to restaurant
6.30pm	Dinner at Brockton Villa, 1235 Coast Blvd, tel. 858-945-4724 (Contact point: Linda Lanyon)

DAY 2 : REQUIREMENTS & IMPLEMENTATION ROADMAP

Tuesday 14th Jan	**Room 3004, Atkinson Hall Building**
8.30am	Gather in hotel lobby to walk to UCSD
8.30am	Breakfast available in meeting room
9-9.15am	Meeting begins: aims for today Maryann Martone
9.15-1030am	Use Cases (brainstorming session, moderator: Sean Hill) Chinh Dang, Allen Institute Use Cases Stephen Larson, UCSD Everyone: brainstorming use cases
10.30-10.45am	Coffee Break
10.45am-1230	Requirements (brainstorming session, moderator: Maryann Martone) • Prioritising and satisfying use cases • Which parts of encyclopedia to be populated first
12.30-1.30pm	Lunch
1.30-3pm	Implementation (moderator Sean Hill): • a model of cooperation, coordination & collaboration • what resources are able to be contributed (services, platforms, funding) • roadmap: a feasible timeline
3-3.30pm	Coffee Break
3.30-4.30pm	Summary & next steps: plan for further meetings – discussion led by Maryann Martone
6pm	Dinner at Mustang & Burros at Estancia hotel for those people staying (Contact point: Linda Lanyon)

Neuroscience

INCF Secretariat
Karolinska Institutet
Nobels väg 15 A,
Stockholm, 171 77 Sweden

Tel: +46 8 524 870 93
Fax: +46 8 524 870 94
E-mail: info@incf.org
Web: www.incf.org