1st INCF Workshop on Cyberinfrastructure for Neuroinformatics

December 9th, 2011
Stockholm, Sweden
1st INCF Workshop on Cyberinfrastructure for Neuroinformatics

Goal

INCF recently started efforts to establish a cyberinfrastructure for neuroinformatics, including federated filesystems and object models for sharing neuroscientific data, and standards for workflows. The goal of this workshop is to discuss the development of an international federated dataspace for large-scale heterogeneous data.

Date and Location

December 9, 2011
Hall Strix, Von Eulers väg 4, Karolinska Institutet, Stockholm, Sweden

Authors

Satrajit Ghosh, MIT, Cambridge, MA, USA
Christian Haselgrove, U Mass, Worcester, MA, USA

Scientific Organiser

Erwin Laure
PDC
Royal Institute of Technology (KTH)

Workshop Participants

Ilkay Altintas, SDSC, San Diego, CA, USA
Nenad Buncic, EPFL, Lausanne, Switzerland
Predrag Buncic, CERN, Geneva, Switzerland
Chinh Dang, Allen Institute, Seattle, WA, USA
Andrew Davison, CNRS-UNIC, Gif-Sur-Yvette, France
Mark Ellisman, UCSD, San Diego, CA, USA
Satrajit (Satra) Ghosh, MIT, Cambridge, MA, USA
Jeff Grethe, UCSD, San Diego, CA, USA
Christian Haselgrove, U Mass, Worcester, MA, USA
Sean Hill, INCF Secretariat, Stockholm, Sweden
Kay Inagaki, RIKEN BSI, Wako-Shi, Japan
Peter Kunszt, ETH Zürich, Zürich, Switzerland
Stephen Larson, UCSD, San Diego, CA, USA
Erwin Laure, KTH, Stockholm, Sweden
Paul Millar, DESY, Hamburg, Germany
Reagan Moore, RENCI, Chapel Hill, NC, USA
Raphael Ritz, INCF Secretariat, Stockholm, Sweden
Ola Spjuth, KI, Stockholm, Sweden
Ilya Zaslavsky, SDSC, San Diego, CA, USA
Tiziano Zito, TU Berlin, Berlin, Germany
Per Öster, CSC, Espoo, Finland

Supported by the INCF and Karolinska Institutet.
## Contents

1st INCF Workshop on Cyberinfrastructure for Neuroinformatics

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Summary</td>
<td>4</td>
</tr>
<tr>
<td>Introduction</td>
<td>4</td>
</tr>
<tr>
<td>Concepts</td>
<td>5</td>
</tr>
<tr>
<td>Workshop Discussions</td>
<td>6</td>
</tr>
<tr>
<td>1. Requirements and Challenges</td>
<td>6</td>
</tr>
<tr>
<td>2. Use Cases</td>
<td>7</td>
</tr>
<tr>
<td>3. Proposed Solutions/ Technologies</td>
<td>7</td>
</tr>
<tr>
<td>4. Technical Requirements and Layout</td>
<td>8</td>
</tr>
<tr>
<td>5. Recommendations</td>
<td>8</td>
</tr>
<tr>
<td>Phase I: Foundations</td>
<td>8</td>
</tr>
<tr>
<td>Phase II: Iterative Building</td>
<td>9</td>
</tr>
<tr>
<td>Appendix A: Workshop Program</td>
<td>10</td>
</tr>
</tbody>
</table>
Executive Summary

The goal of a global computational infrastructure for neuroscience is becoming increasingly important as data sharing and integration grows in importance and data sets and analysis grow in scale and complexity. The scale of the problem makes it intractable to solve at once, and smaller scale solutions have appeared to address subsets of the problem in subdomains of neuroscience. This makes any effort to reach an integrative solution more important; indeed, it is vital that we address this now to align and coordinate efforts. The purpose of this workshop was to discuss high-level requirements and technical needs for a global neuroinformatics infrastructure and to set down a road map for its implementation.

Although an agenda was proposed for initial structure to the workshop, it provided only loose guidance in which ideas could flow freely: high-level requirements were discussed both in general and in the context of the Digital Atlasing Program’s efforts, and existing solutions to the general problem (iRODS, dCACHE, EUDAT) were presented. It was agreed to move ahead with an infrastructure for data publication and sharing that supports searching and discovery, working across institutions and subfields. User authentication and authorization must be supported, preferably using a single sign-on system. Ultimately it should be as easy to use as a user’s familiar personal machine, and support existing solutions (databases and infrastructures). While the system is distributed as much as possible, the INCF can provide technical coordination.

The proposed road map for actual development is broken into two parts: building the foundation of the system and iteratively building functionality on top of that. Setting up the foundation will require securing node resources, ensuring connections between nodes, and setting up management systems, including testing and monitoring environments. Once this is in place, a basic services stack will be built on top of the basic infrastructure to support cataloging, discovery, analysis, automatic metadata extraction, and visualization, and use cases will be developed to validate the system, and maintenance will be addressed, including monitoring, auditing, governance model.

Introduction

Neuroscientific data, associated computing technologies and collaborations are increasing at a tremendous rate. Significant progress has been made in acquiring data at multiple spatial and temporal scales of neuroscience with a myriad of acquisition technologies. Furthermore, current social networking technologies have addressed various authentication and authorization issues and peer-to-peer and distributed storage technologies have addressed file distribution issues. Yet, the ability to catalog and share and the access (especially public access) to neuroscientific data have remained limited. Furthermore, even when such data are available, relatively few organizations and laboratories have the resources to perform computation on such data. Therefore, even with the extensive growth of computational technologies over the last decade, several barriers remain in creating an infrastructure that allows efficient access to neuroscience data and associated computation capabilities.

The goal of this workshop was to:

1. Highlight current barriers that limit sharing, querying, accessing and operating on data from different laboratories
2. Evaluate solutions that overcome these limitations
3. Recommend a roadmap for an infrastructure that can be built to accommodate data distribution and querying needs
### Concepts

<table>
<thead>
<tr>
<th><strong>Computational Infrastructure/Federated Infrastructure/Cyber-infrastructure:</strong></th>
<th>A technology infrastructure comprising compute nodes and storage locations for world-wide storing, sharing, querying, and analyzing data.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Authentication, Authorization</strong></td>
<td>Authentication is the act of verifying that a user is who they say they are. Authorization is granting a user certain permissions. Both are required for a working security model.</td>
</tr>
<tr>
<td><strong>Ontology</strong></td>
<td>A structured domain specific vocabulary including precise definitions of and relationships between terms.</td>
</tr>
<tr>
<td><strong>Scalability</strong></td>
<td>The ability of a system to maintain its performance (e.g. query response times, processing durations, download throughput) when applied to more sites, more users, more data, or larger data sets.</td>
</tr>
<tr>
<td><strong>Grid</strong></td>
<td>Several computers working together, either for processing (a computational grid) or storage (a storage grid).</td>
</tr>
<tr>
<td><strong>Cloud</strong></td>
<td>Computational or storage resources that are not local to the user. These may be distributed, but one defining characteristic is that the user does not need to know the details of how the cloud is physically constructed.</td>
</tr>
<tr>
<td><strong>Catalog</strong></td>
<td>Standard metadata across data sets that is brought together to facilitate data access and discovery.</td>
</tr>
<tr>
<td><strong>Governance</strong></td>
<td>The process of active management of the direction of a project. This will often include continued monitoring of the project and use of its systems and refinements or redefinitions of requirements. Governance typically involves a broad base of stakeholders.</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>Data about or associated with data used to render a more precise description or record of its significance. An item of metadata may describe an individual data item or a collection of data items and is used to facilitate the understanding, use and management of data.</td>
</tr>
<tr>
<td><strong>Federate</strong></td>
<td>To make separate resources work together as a single unit. For example, data shared in separate databases may be joined so a single query can expose data across these databases. (This would be a federated data space and a federated query.) A grid results from federation of computational resources.</td>
</tr>
<tr>
<td><strong>Publish</strong></td>
<td>To make data publicly available. This need not be publication in the traditional sense of journals and articles; indeed, in this context, publication is more likely to refer to simple accessibility like being available on a web site.</td>
</tr>
<tr>
<td><strong>Share</strong></td>
<td>To make data available to others, either publicly or privately.</td>
</tr>
<tr>
<td><strong>Annotation</strong></td>
<td>A comment on data that lives with the data (in other words, not a comment in the sense of a journal letter). Annotations may be public or private, and may refer to a very specific part of the data or a larger subset of the whole data set.</td>
</tr>
<tr>
<td><strong>Workflow (or pipeline)</strong></td>
<td>A description of steps performed during data analysis, which captures dependencies between processing steps and defines parameters that can be set during the analysis. Typically a workflow is expressed in a machine-readable form, facilitating immediate use by a workflow or pipeline engine, which performs the analysis. Structured workflows facilitate provenance tracking, reproducibility, analysis standardization, and communication of methods.</td>
</tr>
<tr>
<td><strong>Node</strong></td>
<td>A national unit under the INCF umbrella. Loosely speaking, a node refers to a country's participants in INCF activities and their projects and resources that support INCF activities.</td>
</tr>
</tbody>
</table>
Workshop Discussions

Being an unconference and following open space principles, discussions followed only a loosely structured agenda (see below) and were allowed to meander, and the result was disparate discussions from which unstructured requirements came. In this document, we summarize presentations and discussions, highlighting challenges, requirements, and solutions that are addressed in the Recommendations section below.

General outcomes of talks and discussions fall into three major categories:

1. Requirements and Challenges
2. Use Cases
3. Proposed Solutions/Technologies

1. Requirements and Challenges

Sean Hill started the workshop with background and an overview of the problem. He introduced the INCF and its programs, and described existing INCF efforts for a neuroinformatics infrastructure. He described a vision for a layered infrastructure model that would build an object model on top of a federated data space and higher-level functionality on top of that. Key to the workshop was his assertion that “to share is not enough”; a federated infrastructure must allow for:

• Publishing data, making it citable, and allowing the community to comment on it
• Preserving data, including support for versioning and archiving
• Federating data across tens to hundreds of thousands of users
• Searching across all data, using free-text searches or structured searches using an ontology or spatial concepts
• Accessing data, its annotations, metadata, and models, possibly in an object-based manner
• Analyzing and integrating data, supporting workflows and provenance tracking
• Ease of use of all of the above

Ilya Zaslavsky then reported on the INCF Digital Atlasing Infrastructure. The Digital Atlasing Program is the most mature of the INCF Programs requiring a computational infrastructure, so this presentation provided a detailed description of a domain that the cyberinfrastructure in question should support, including implicit requirements and use cases as well as current challenges.

In broad strokes (details are described in detail elsewhere: http://atlasing.incf.org/wiki/Main_Page), the Digital Atlasing Program’s efforts are concerned with the analysis of rodent brain images using atlases. A standard space (Waxhom Space) was defined, and end users can register their images to this space and query other resources available in this space. A range of issues come into play, including but not limited to a standard language for communication (WaxML), definition of web services for operations, data and metadata representation, processing workflows and provenance, and data sharing and annotation.

The Digital Atlasing requirements for the cyberinfrastructure are:

• Support for publishing, cataloging, discovery and access to key types of neuroscience data
• Enabling straightforward sharing of large datasets via common elastic virtual data space (data grid, cloud)
• Enabling distributed computation and effective pooling of compute resources (computational grid)
• Common and transparent information model and API foundation that developers can use to contribute code, applications, data and resources
• Scalability to additional users, larger datasets, new types of data and resources
• Support for long-term preservation and on-demand availability of neuroscience resources of different types
• Support for a range of integration models: ability to integrate data across scales; species; development stages
• Ability to compose and run complex workflows, and ensure provenance tracking, re-usability and accuracy assessment of the results
• Enabling re-use of data by adopting or developing key interoperability standards and reference frameworks, including semantic and spatial reference models
• A governance model (rather, models) that addresses expectations of all stakeholders

Further discussion identified additional issues that should be addressed:

• Identifying which aspects of the system are critical for standardization. In this case, catalogs, vocabularies, process descriptions, and data model commonality (of some sort) were identified
• Deciding on the right level of abstraction and scoping using use cases
• Reproducibility and unambiguous interpretation
• Identifying the subset of the complete solution that will best move the community in the right direction
• Identifying sources of error and representing error
• Supporting public (third-party) and private annotations of data
• Cost models

Finally, it was noted that it will be essential to keep in mind the motivating factors for the cyberinfrastructure for both researchers and funding agencies.

2. Use Cases

Some of the use cases that the iRODS-based INCF Cyberinfrastructure is intended to support include:

- Allowing researchers to disseminate data to other scientists in a globally distributed fashion, as stipulated by some grants, whether or not the other scientists are members of INCF. The INCF Cyberinfrastructure must make this as easy as possible for the researcher. The researcher in this case might be part of an organization who already has a data resource being shared, where they can place their data to be published, but the infrastructure also needs to be suitable for researchers who might need to join an existing INCF regional node and provide their own storage, all the way to the case where the researcher needs to integrate with the infrastructure by creating their own iRODS zone.

- Providing a secure, reliable and searchable archive for experimental and analysis data. Not only can the INCF Cyberinfrastructure be used as a “backup” for important data, but data can also be tracked through its lifecycle, even though individual researchers might come and go from a particular project, research group or organization. No longer does a PI need to worry that the only copy of an important dataset disappears on the laptop of a leaving postdoc, or is a casualty of a corrupted hard drive, since there are many options for data storage and protection within the Infrastructure.

- Browsing of very large datasets in place, regardless of whether they are located within a researcher’s organization, or across the world in the data repository of another INCF collaborator. One can imagine a researcher browsing thumbnails of very large files remotely, so that they only need to retrieve the subset of data actually needed. For some types of files, in place browsing could also include retrieving single slices of data (e.g. parts of a BAM file representing genes of interest).

- Having a mechanism for annotating data with meta-data, perhaps using automated extraction, and allowing users to search the entire INCF Cyberinfrastructure using these meta-data terms. The infrastructure could also allow researchers to apply well-known and/or standardized ontologies as they become more widely defined and used.

3. Proposed Solutions/Technologies

iRODS

Reagan Moore gave a presentation on iRODS, a policy-based data management system. iRODS is a layered system that supports storage of data, data access control through policy enforcement, and high-level interfaces for data access. iRODS supports metadata cataloging and searching, and is scalable and can be integrated across sites. It is open source, in use supporting a number of other efforts, and has several client-side tools available to ease access (e.g. iDROP, which facilitates adding data to an iRODS grid). Features under development that were of interest to the workshop include pluggable authentication modules and workflow support.

dCache

Paul Millar described dCache, a storage system now
used in the physics community for over a decade. At the time of the workshop, 383 sites were federating 101 PiB. DCache is standards compliant (HTTP, WebDAV, FTP, NFS), supports pluggable authentication, and features easy integration with tape system.

Pros of dCache are its standards-based architecture, but some local access protocols are proprietary and the data flow is based on an outdated model.

Current efforts include work on an HTTP redirector that will allow HTTP-based access to data for WAN users and NFS access for LAN users, providing ease of use by using existing standards to allowing users to use unmodified client software.

EUDAT

Per Öster spoke on EUDAT, a European collaborative data infrastructure. EUDAT is a collaboration of 25 partners (data centers, research institutes, and tech providers) in 13 European countries which currently serves several disciplines including geology, climate science, and physiology. The initial EUDAT schedule runs for three years and is in its first year, during which requirements are gathered, challenges are identified, and initial services are implemented.

The goal of EUDAT is to provide a generic data infrastructure that is flexible, sustainable, and spans geographic and disciplinary boundaries. The challenge of spanning various countries, fields, and existing technologies has led EUDAT to the solution of interoperability of resources (rather than resource integration or trying to ensure compatibility or adaptability), which will scale better given the challenges at hand.

The proper role of centralization has also been carefully considered, and while centralization is crucial, it must also be limited in order to accommodate political and legal boundaries and existing disciplinary and local (geographic) infrastructures.

4. Technical Requirements and Layout

A discussion of technical requirements and layout served to marshal requirements and ideas and lay the stage for concrete recommendations for moving forward.

The goal of the system is for an international infrastructure that spans various countries, in several subdomains, and supports several technical systems currently in use. A data sharing and querying infrastructure is the primary target, allowing users to easily add their data to the system, attach permissions to the data, and allow for other users to access the data. Being able to search for and discover data is of great importance. An important trait of the system will be its ease of use; an idealized goal is to be able to interface with the infrastructure in a Dropbox-like manner, fully integrating with a user’s local machine.

Technical discussions suggested implementing a data layer on top of existing local storage, domain databases such as XNAT, HID, IDA, and CCDB, and existing national infrastructures. The data layer would provide for registration and cataloging of data and metatdata and might provide a persistent ID for data. iRODS could play a central role in this data layer, and the INCF could provide a coordinating function without needing to be a central repository for data.

User authentication and authorization is another important consideration, and not trivial to integrate across existing systems. Implementing a single sign-on approach may be the best way to go: this is common and relatively easy to implement for a local solution, and the INCF could provide a coordinating role, including providing accounts for users who do not have a local system that will provide an account for them.

Additional considerations included the importance of identifying and supporting common terminologies and being able to support analysis and workflows. While neither of these are explicitly addressed in the recommendations below, they are important functions that should be supported in the future.

5. Recommendations

The workshop participants recommended moving forward on a technical infrastructure to meet the needs discussed in the workshop and described above. Specific actions were outlined in two phases: laying down the building blocks for the infrastructure and iteratively building on top of them.

Several national nodes were identified to participate in these development activities:

- Finland
- Sweden
- Norway
- United States
- Switzerland
- Germany
- France (in consultation with IN2P3)

Phase I: Foundations

The first phase will be setting up the foundations for the infrastructure. Resources will be secured and low-level systems as well as some basic higher-level management functionality will be put into place. Authentication will be done nationally by Shibboleth or SAML systems, and users without access to one of these systems locally may register centrally for authentication with the INCF. The INCF will also maintain lists of groups of users.
1. Secure node resources. Approach national infrastructures to request storage and a single sign-on system based on an iRODS back end to federate national nodes. Set up a central iRODS instance at the INCF to handle authentication for users without local Shibboleth or SAML systems and for managing groups. RENCI will provide a procedure to distribute the group information from the central INCF server to local iRODS instances.

2. Inter-node connection. Federate the national nodes’ iRODS infrastructures between nodes. The Finnish and Swedish nodes have experience with this.

3. Management systems. Make sure iDrop is working with metadata catalogs. Set up testing and monitoring environments.

**Phase II: Iterative Building**

The second phase will build on top of the basic infrastructure to apply it to practical uses. An analysis use case as well as integration of certain existing systems are the targets for this phase. Planning for long-term sustainability will then begin.

1. Build the stack. Add services on top of the basic infrastructure to support cataloging, discovery, analysis, automatic metadata extraction, and visualization. Specific services to be integrated are:
   - Digital Atlasing infrastructure
   - Neuroscience Information Framework
   - INCF Brain Catalog

2. Analysis use case. Develop and test a simple analysis use case. At this stage the use case will be limited; more complex workflows will follow later.

3. Develop an iRODS instance to function as a “portable CCDB.” This could be started as part of Step 3: Management Systems above.

Appendix: Workshop Program

Thursday, December 8, 2011
18:30  Dinner at Oljebaren (Vasastan)

Friday, December 9, 2012
9:00 - 9:45  Introducing the INCF
            Sean Hill
9:45-10:30  The INCF Digital Atlasing Infrastructure - status & plan
            Ilya Zaslavsky
10:30-11:00 Coffee break
11:00-12:30 Session I: Building an international distributed system for data sharing- Requirements & Layout
            Chair: Jeff Grethe
12:30-13:30 Lunch
13:30-15:00 Session II: User Experience & Expectations: drill-down on use cases
            Chair: Mark Ellisman
15:00-15:30 Coffee break
15:30-17:00 Session III: Refine/recommend actions/timeline
            Chair: Sean Hill
19:00  Dinner