**NCANDA: Data Analysis Component**

**Objective**: Alcohol and marijuana remain the most commonly used central nervous system active substances in teen years. The National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA) maintains a data repository of longitudinal, quantitative brain imaging and cognitive assessments designed to capture the influence of adolescent alcohol abuse on neurodevelopment. The goal of this proposal is to enhance the long-term sustainability of NCANDA's neuroimaging and cognitive data by extending and incorporating data exchange standards into the repository.

**Problem**: Heterogeneous data model semantics are inherent to complex longitudinal study protocols. The diversity of data collected by these studies requires biomedical data management and electronic data capture systems tailored to specific use-cases. These study-specific solutions generally result in data silos making data integration difficult. For example, the parent grant of this supplement has been focusing on developing a software platform for harmonizing the multi-modal data collected within the NCANDA consortium. Easing interoperability with the scientific community is outside their scope of work; a serious limitation to the accessibility and future use of the 20 Terabyte dataset created by NCANDA.

**Proposed Solution**: The four sites of this collaboration will create a data exchange standard that enriches neuroimaging studies with structured data for aiding information retrieval, data integration and archiving. Using NCANDA (SRI) as a driving application, we will augment the primary data with a data exchange specification framework, called the Neuroimaging Data Model or NIDM standard (Berkeley), and the Cognitive Atlas ontology (Stanford). The resulting metadata will be ingested into the Synapse data repository (Sage Bionetworks) in order to demonstrate interoperability with the NCANDA dataset. The additional specific aims to the U01 entitled "NCANDA: Data Analysis Component" proposed in this supplement are as follows:

**Aim 6**: Extend Data Exchange Specifications of NIDM with Brain Imaging and Cognitive Measures of NCANDA
**Approach:** We will use a collaborative design process to enhance NIDM, a structured data exchange standard, with the Cognitive Atlas ontology. We will do so by first extending the Cognitive Atlas ontology with new terms motivated by NCANDA data. We will then review the resulting data model with the data sharing and neuroinformatics community to account for their needs in the design.
**Hypothesis:** The resulting structured data exchange will enable interoperability between the NCANDA and Synapse repository.
**Evaluation:** We will test this hypothesis by identifying information retrieval tasks that the data exchange specifications support. We will conform a subset of the NCANDA data to complete these tasks. The team at SRI will load the conformed data into the Synapse repository. Using the Synapse Web Application Programming Interface (API), Sage Bionetworks will independently validate each information retrieval task.

**Aim 7**: Embed Data Exchange Standard into the NCANDA Software Platform
**Approach**: We will extend the NCANDA software platform to interoperate with the Synapse repository based on the data exchange specifications developed in Aim 1. The software first conforms the NCANDA data to the data exchange specifications. It then accesses a Web Service API to map neuropsychological data to the Cognitive Atlas ontology. The resulting semantically annotated data will then populate the Synapse repository using the Synapse API. We will unit test the software extension to ensure its proper operation and further check correct storing of the primary data of NCANDA on Synapse.
**Hypothesis:** The proposed extension to NCANDA will enhance the query capabilities of the repository.
**Evaluation**: Based on the ontology-driven metadata added to the NCANDA data, SRI will create a set of queries impossible to perform on the primary data such as "finding the dimensions of neuropsychological function related to substance use and their relation to brain shape measures". Sage Bionetworks will validate those queries on Synapse.

**Impact**: This project will establish a new approach using a neuroimaging data exchange standard that can be reused and extended by the broader community. Applied to the NCANDA repository, this data exchange standard will increase the interoperability of this dataset by providing machine-readable metadata that is portable to heterogeneous data repositories, such as Synapse. Semantic annotation will result in an extensible querying system across distributed resources with the potential of identifying new findings on the effect of alcohol or drug abuse on adolescent neurodevelopment based on large scale meta-analysis.

## RESEARCH STRATEGY

### A. Significance

#### A.1. Importance of the problem

Alcohol and marijuana remain the most commonly used central nervous system-active substances in the teen years [1]. The National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA) is a multisite, longitudinal "Big Data" study using quantitative assessment tools necessary to capture the influence of adolescent alcohol and marijuana abuse on neurodevelopment. The goal of this proposal is to 1) ensure sustainable long-term access to shared primary data (~6TB) and computed results (~20TB) produced by the NCANDA data repository [2] and 2) develop data exchange standards and software that will facilitate interoperability between neuroimaging data repositories [3]-[6], specifically with respect to neurodevelopmental measures sensitive to Alcohol Use Disorder (AUD) and Substance Use Disorder (SUD).

#### A.2. Technical barriers

Heterogeneous data models and semantics are inherent to complex study protocols that capture rich neuroimaging and neuropsychological measures [7], [8]. The diversity of data collected by these studies requires biomedical data management and electronic data capture systems tailored to specific use-cases [9]-[13]. For example, the successful operation of the NCANDA study required the development of a data integration system [2] to merge the multi modal data from different repositories (i.e. WebCNP [14], REDCap [15], XNAT [16]). These study-specific solutions generally result in data silos and stovepipes that limit an integrated view of the information outside the original scope of analysis [17], [18]. Incorporating metadata standards in the design of these systems can streamline data integration processes and interoperability, facilitating submission to national data repositories and querying data across studies. The use of metadata standards is currently outside the scope of most medical imaging studies, such as NCANDA, which is a serious limitation to the current and future accessibility and use of their data.

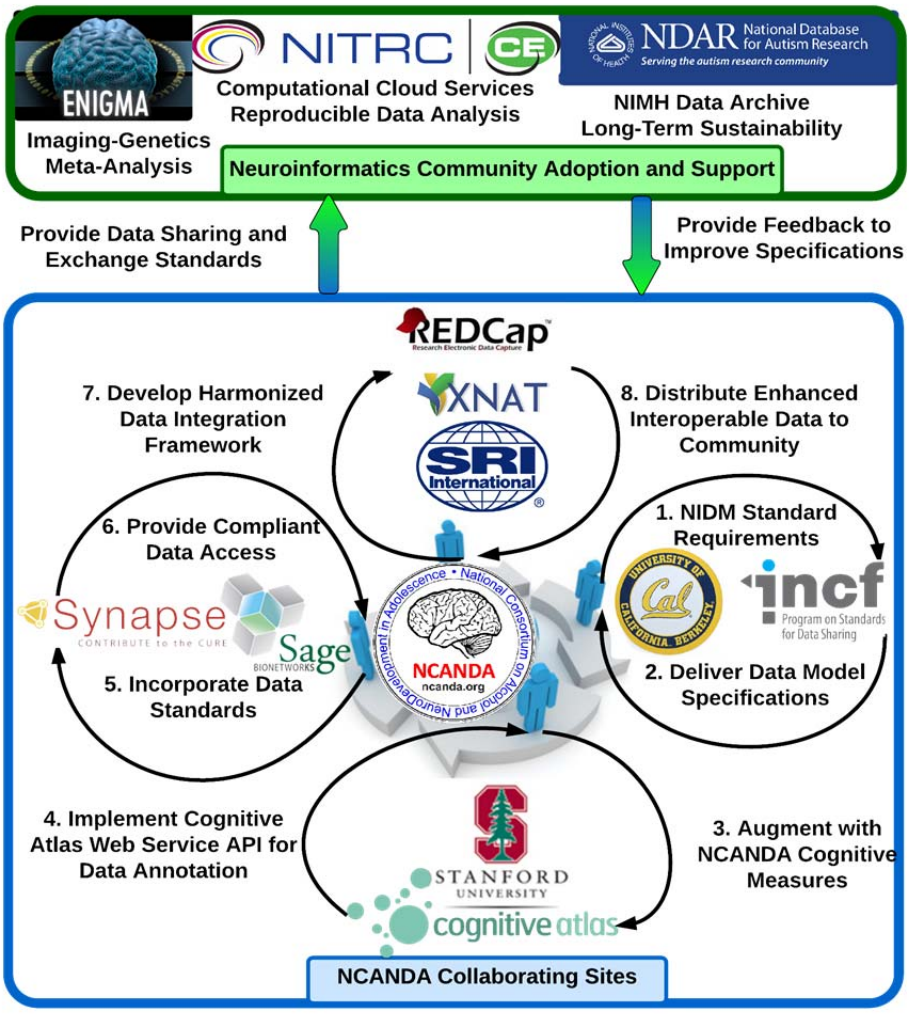


**Figure 1.** Overview of activities completed by each of the 4 collaborating sites and impact on broader community.

#### A.3. Impact of the proposed project

In this project (Figure 1), we will extend the existing NCANDA infrastructure [2] with technologies designed for data interoperability to **a)** augment primary data (i.e., neuroimaging, cognitive measures) from NCANDA (SRI)

using **b)** the community-driven data exchange approach defined by the Neuroimaging Data Model (NIDM, [19]; UC Berkeley) with **c)** semantic annotations from the Cognitive Atlas ontology (Stanford) and **d)** apply the data exchange standards to interoperate with the Synapse Web Service API (Sage Bionetworks). The resulting data integration framework will enhance an emerging standard for neuroimaging data exchange that can be reused and extended by the broader community. By enhancing NIDM with additional neuroimaging and cognitive measures we will also increase the interoperability of the NCANDA data with machine-readable metadata portable to heterogeneous data repositories, such as Synapse. In other words, the resulting semantic annotations create an extensible querying system across distributed resources with the potential of identifying new findings on the effect of alcohol or drug abuse on adolescent neurodevelopment based on large scale meta-analysis.

### A.4. Responses to Specific Requirements of PAR-15-144

In the following sections we respond directly to each of the research strategy requirements listed in Section IV.2 PHS 398 Research Plan of the FOA.

#### A.4.1. Description, Role, and Eligibility of Collaborating Sites

**SRI International (SRI)** is a nonprofit institute with 501(c)(3) IRS status that was awarded the active NIH funded parent grant entitled "NCANDA: Data Analysis Component" (5U01AA021697). The Data Analysis Component is currently creating a biomedical data repository that supports the storage and distribution of brain imaging, biological samples, clinical, and cognitive data collected from the five NCANDA consortium sites (see Table 1). Drs. Pohl and Nichols (Key Personal) will coordinate the design of data exchange specifications between collaborating sites and oversee the implementation of streamlining data sharing from the NCANDA repository to Synapse and possible other repositories (see letters of support by Dr. Thompson for Enigma, Dr. Kennedy for NITRC, and Mr. Hall for NDAR) based on interoperability standards.

**University of California, Berkeley (Berkeley)** is a public institution of higher education that is the recipient of many NIH grants. Dr. Poline (key personnel, see letter of support) will oversee the data specification efforts of the proposed project with a standardized data directory structure and metadata exchange effort (NIDM) initiated by the International Neuroinformatics Coordinating Facility (INCF). Seventeen member countries fund INCF, including the NIH and NSF for the USA.

**Stanford University (Stanford)** is a private institute of Higher Education that was previously awarded the NIH grant "The Cognitive Atlas: Developing an Interdisciplinary Knowledge Base" (5R01MH082795). The Cognitive Atlas primarily supports a community-driven biomedical ontology that relates psychological functions to brain systems. Dr. Poldrack (key personnel, see letter of support), will oversee the implementation of a Web API for the Cognitive Atlas, which maps neuropsychological data of NCANDA to the Cognitive Atlas ontology. The resulting semantically annotated data will then populate the Synapse repository.

**Sage Bionetworks (Sage)** is a nonprofit institute with 501(c)(3) IRS status that was awarded the active NIH grant entitled "Integrating Cancer Dataset for Predictive Model Development and Training" (5U54CA149237) for supporting the collaborative analytics and data sharing environment called Synapse (https://www.synapse.org/). Dr. Klein (key personnel, see letter of support), will lead in integrating the NCANDA metadata into the Synapse repository and aid in evaluating the NCANDA system via use-case specific queries.

| Project No. | PI | Organization | Role | Funding | FY Total |
|---|---|---|---|---|---|
| 5U01AA021697 | POHL, K. M. | SRI INTNL. | Data Analysis | NIAAA | $772,199 |
| 5U01AA021695 | BROWN, S. | UC SAN DIEGO | Administrative | NIAAA<br>NIMH | $25,721<br>$500,002 |
| 5U01AA021692 | TAPERT, S. | UC SAN DIEGO | Data Collection | NIAAA | $798,422 |
| 5U01AA021696 | COLRAIN, I.M. | SRI INTNL. | Data Collection | NIAAA<br>NICHD | $570,445<br>$300,000 |
| 5U01AA021690 | CLARK, D. | U. PITTSBURGH | Data Collection | NIAAA | $509,870 |
| 5U01AA021681 | DE BELLIS, M. | DUKE UNIV. | Data Collection | NIAAA<br>NIDA | $425,272<br>$291,000 |
| 5U01AA021691 | NAGEL, B. | OHSU | Data Collection | NIAAA | $557,174 |
| | | | | **TOTAL** | **$4,750,105** |

**Table 1.** Breakdown of 2014 NIH Funding of NCANDA consortium, which consists of $500K from.NIMH, $300K from NICHD, $291K from NIDA and $3.6M from NIAAA.

**A.4.2. Eligibility of the Parent Grant** The active NIH funded parent grant is entitled "NCANDA: Data Analysis Component" receives an overall annual budget of $594,572 in Direct Costs. The NCANDA Data Component

primarily supports a biomedical data repository of brain imaging, biological samples, clinical, and cognitive data collected from the five NCANDA data collection sites (see Table 1).

### A.4.3. How Successful Completion of the Proposed Project Will Impact Data Interoperability

We will deliver reusable informatics products that make data and knowledge exchange more efficient for neuroimaging studies. We will develop informatics methods for data interoperability between two repositories of biomedical data using a streamlined data integration and interoperability framework that adopts a grass-roots based data exchange standard (NIDM), a vocabulary/ontology-based data exchange standard (Connective Atlas), and a Cloud-based data storage and sharing infrastructure (Synapse) with a Web Application Programming Interfaces (API).

**SRI:** We will extend our existing data integration framework with external interoperability approaches. Specifically, we will incorporate the Cognitive Atlas and NIDM for semantic data annotation of the NCANDA participant records. Further, our data integration framework will leverage the Synapse for Cloud based storage and analytics.

**UC Berkeley:** We will remove barriers to standardized neuroimaging data exchange by augmenting the existing NIDM vocabulary/specifications to support the Open Brain Imaging Directory Structure (OBIDS) being developed at Berkeley in collaboration with the neuroinformatics community. NIDM will be implemented in Synapse to demonstrate the NCANDA data integration approach.

**Stanford:** We will simplify the reuse of machine-readable knowledge about cognitive function by creating a Web Service API to access the concepts, tasks, and conditions contained in the Cognitive Atlas ontology. We will specifically seek to identify requirements and implement functionality necessary for reuse of this ontology in the Synapse repository.

**Sage:** We will remove barriers to scalable data storage and sharing for the NCANDA data repository by developing an automated data interoperability framework that builds on the Synapse repository APIs using open-source, data management software targeted towards brain imaging data studies.

### A.4.4. Evidence the Proposed Project Improves Interoperability Beyond Current Efforts

We will design and implement a data exchange standard that enriches neuroimaging studies with structured data aiding information retrieval, data integration and archiving. Using NCANDA (SRI) as a driving application, we will enhance the NIDM standard (Berkeley) with neuroimaging and cognitive measures, implement an API to the Cognitive Atlas ontology (Stanford) for data annotation, and then populate the Synapse repository (Sage) with NCANDA data and NIDM metadata. We now review the current efforts and proposed advances in interoperability for each site:

**SRI:** While the NCANDA data repository is responsible for sharing its data with the neuroscience community, all harmonization efforts are focused within the consortium itself (aims of parent proposal are listed in the Introduction). The proposed project would extend interoperability to the neuroinformatics community at large.

**UC Berkeley:** The NIDM data exchange standard, an effort led by Berkeley and in partnership with the INCF, is currently focused on interoperable reporting of fMRI results. The proposed project will provide necessary resources to extend NIDM to support the Synapse API, semantic annotations from Cognitive Atlas, and cognitive measures from NCANDA.

**Stanford:** While the Cognitive Atlas ontology has an active user base, further development is currently not supported by outside funding. The proposed API would provide computational access to cognitive concepts and measures captured by the ontology.

**Sage:** Synapse provides a collaborative environment for biomedical investigations. The proposed project would enhance the Synapse with data exchange standards for neuroimaging and cognitive studies.

### A.4.5. Sustainability Plan for How Proposed Activities Will Support and Improve Repositories

The NCANDA data repository contains approximately 850 participants with three time points (baseline, 1-year follow-up, and 2-year follow-up). For each time point, imaging and cognitive tests are collected that incur a data storage and access burden for long term sustainability of this repository. The current clinical data management system is 42GB, the primary (raw) imaging data is 2.7TB, and the pre-processed imaging data is 7.2TB. The size of the final data repository size is estimated around 20TB; however, simply storing the data is not enough. Unless the data is properly curated with metadata by the original investigations it will have minimal utility for researchers outside the consortium, as discovery generally requires integral knowledge about the data [20]. To improve long-term reuse and efficiency of this repository, the proposed activities will deliver a data exchange

standard that can be used across the neuroimaging and cognitive measures enterprise to migrate data into repositories that function as final archival systems with appropriate metadata. The semantically annotated framework will also enable an extensible querying system that leverages Web standards designed to access distributed resources as a web of data. Scientists with access to this web of data have greater potential to identify new findings on the effect of alcohol or drug abuse on adolescent neurodevelopment compared to traditional data management approaches.

## B. Innovation

### B.1. Challenges to Current Paradigms

The current data sharing paradigm consists of heterogeneous data repositories [10], [12], [21]-[23]. These repositories have grown in isolation from one another other and are commonly referred to as data "silos" or "stovepipes" due to their lack of cross communication. Similarly, NCANDA developed its own internal data harmonization approach that is incompatible with external data repositories [2]. The semantics used to model information varies considerably across data sources making data integration processes a daunting and time consuming task [17]. In summary, the lack of standard data exchange protocols severely limits the ability to automate data integration tasks [24].

### B.2. Novel Concepts, Approaches, and Methodologies

Semantic Web and Linked Open Data technologies have matured since their conception [25] and are actively developed in many biomedical domains [26]-[31]. For example, in molecular biology and genomics, semantic Web technologies have proven their usefulness to facilitate interoperability [32]-[34], and in industry (e.g., Google's Knowledge Graph) they have demonstrated their scalability. However, there are only a few examples of software applications that apply these technologies to human brain imaging [19], [35], [36] and cognitive measures [4]. In this proposal, we advocate the application of semantic Web principles to extend NIDM, an emerging neuroimaging data exchange standard, with enhancements to represent brain imaging and cognitive measures in a model supporting advanced information retrieval. For example, researchers will be able to expand their queries using ontology terms rather than using keywords, thus enabling information retrieval tasks that take advantage of knowledge encoded in ontologies.

### B.3. Improvements over Existing Methods

In this project, we will improve the state of data exchange in human brain imaging research by using the above technologies to integrate several sources of data, knowledge, and web services. Existing methods do not provide a scalable approach to bridge data models with ontologies in the domain of brain imaging because they lack a mechanism that enables an active community to define and vet a controlled vocabulary in a grassroots fashion. To advance community engagement, we will use an established, open data modeling process that leverages web-based version control system for public discussion and debate (i.e., GitHub, https://github.com/incf-nidash/nidm). Over the past two years, we have found this approach to be very effective for developing NIDM. Specifically, this collaborative framework will enable primary data from NCANDA to be enhanced with semantic annotations from the augmented Cognitive Atlas ontology [4]. The metadata and corresponding annotations will be described with an extended version of the NIDM standard developed by INCF, which organizes brain images using the Open Brain Imaging Directory Structure (OBIDS). This enriched dataset will be ingested into the Synapse repository. The result will be a semantically enriched and standardized representation for neuroimaging data and cognitive measures with specifications made freely available to the community.

## C. Approach

We will extend technologies developed at each of the collaborating sites to demonstrate a semantically enhanced NCANDA data integration framework [2]. As outlined in Figure 1, the NCANDA data repository will be used as a driving use-case for each of the site projects. To develop extensions to NIDM that support primary neuroimaging and cognitive measure, SRI will work collaboratively with Berkeley and in partnership with the neuroinformatics community to draft data exchange specifications (Figure 1-1, 1-2.). SRI and Stanford will cross validate cognitive measures from NCANDA with the Cognitive Atlas to enhance available knowledge linking cognitive processes to tasks (Figure 1-3), while designing and implementing an API to facilitate reuse of this knowledge for semantic data annotation (Figure 1-4). With the data exchange standards in place, SRI will work with Sage to incorporate standardized descriptions of the NCANDA data into the Synapse platform using

their API (Figure 1-5, 1-6). Finally, SRI will finalize development of their semantically enriched data integration framework that ties together components from each of the collaborating sites to provide an open source toolkit for distributing neuroimaging and cognitive data (Figure 1-7, 1-8). An important aspect of this work is to encourage adoption of the data exchange standard and solicit feedback from the broader community (Figure 1: Top), as indicated by the letters of support from Paul Thompson (ENIGMA), David Kennedy (NITRC), and Dan Hall (NDAR/NIMH Data Archive).

**C.1. Aim 6:** Extend Data Exchange Specifications of NIDM with Brain Imaging and Cognitive Measures of NCANDA

**Significance & Innovation:** Heterogeneous data model semantics are inherent to complex longitudinal study protocols. The diversity of data collected by these studies requires biomedical data management and electronic data capture systems tailored to specific use-cases. These study-specific solutions generally result in data silos making data integration difficult, such as in the case of NCANDA. Using NCANDA as a driving application, we will address this issue by extending NIDM, an innovative and open approach to the development of data exchange standards for neuroimaging data. NIDM will provide machine-readable metadata that is portable to heterogeneous data repositories enabling interoperability between the NCANDA and Synapse repository.

**Preliminary Results:** We created the Neuroimaging Data Model (NIDM, http://nidm.nidash.org) that provides a parsimonious description of neuroimaging data provenance using semantic Web technologies [19]. To do so, we used a community-driven and open data modeling process to specify a common shared vocabulary of concepts used in the neuroimaging community and incorporated external terms based on a principle of reuse [19]. We have also developed the Cognitive Atlas as a web application (http://www.cognitiveatlas.org/) that provides a ontology for cognitive concepts, tasks, conditions, and theories [4].

**Approach:** Our approach for both neuroimaging data and cognitive measures is to define standard data exchange specifications that follow from a given use-case. First, we will identify use-cases by gathering requirements from each of the participating sites to identify overlaps in data model and vocabulary semantics. From the identified use cases, we will select a subset from which to design formal specifications (i.e., object models) for data exchange. The object models will be represented using the Web Ontology Language (OWL, [37]). OWL provides a machine-readable format that can be used to generate a human-readable documentation (e.g., see the NIDM Results specification: http://nidm.nidash.org/specs/nidm-results_020.html). We will complete this activity by following a collaborative and community-driven modeling process defined by NIDM using the freely available GitHub platform. This process will be used to curate necessary terms, then design and deliver the data exchange specifications for this project. We will apply this process to both the 1) cognitive measures and 2) neuroimaging data:

**1) Specification of cognitive measures.** We will first generate an inventory of the cognitive measures used by NCANDA. We will then map this inventory to terms of the Cognitive Atlas [4]. Measures not linked to an ontology will be added to the Cognitive Atlas. We will extend the NIDM standard with this ontology. To ensure that the extended NIDM standard is sufficiently general for the neuroinformatics community, we will carefully review the standard with the INCF and other large neurodata repositories (Figure 1: Top), such as ENIGMA (see letter of support from Paul Thompson), NITRC (see letter of support from David Kennedy), and NDAR (see letter of support from Dan Hall).

**2) Specification of neuroimaging dat**a: We will use the NIDM framework to augment the “Open Brain Imaging Data Structure”, a common directory structure developed at Berkeley and the INCF to organize neuroimaging data and facilitate automated analyses. This object model will capture a Resource Description Framework (RDF, [38])-based representation of the directory structure that can be annotated with terms from standard vocabularies or directly curated into NIDM. In collaboration with Sage Bionetworks, we will harmonize this model with the Synapse repository.

**Evaluation:** We will perform a set of information retrieval tasks to ensure that the proposed structured data exchange will enable interoperability between the NCANDA and Synapse repository. Specifically, we will identify information retrieval tasks supported by the specifications. We will then conform a subset of the NCANDA data to complete these tasks. The team at SRI will load the conformed data into the Synapse repository. Sage will execute the retrieval tasks on Synapse while SRI will independently perform the same

queries on the NCANDA repository. We will verify interoperability between Synapse and NCANDA by comparing the query results across repositories.

**Risk and Alternatives:** Designing standards and specifications can be a time consuming process. Should the specification of the data exchange standard risk going beyond 6 month (see also Timeline in Section C.4) then we will confine the standard to a subset of the terms used in the NCANDA data.

**C.2. Aim 7:** Embed Data Exchange Standard into the NCANDA Software Platform
**Significance & Innovation:** The NCANCA data integration framework [2] overcame barriers to interoperability within our consortium by creating a harmonized representation of neuroimaging and cognitive data, yet interoperability with external repositories remains a manual endeavor. For example, metadata about neuroimaging and cognitive measures in NCANDA were developed internally and lack a mapping to the external data models and ontologies used by external repositories. We will now address this issue by embedding the extended NIDM data exchange standard of Aim 1 in the NCANDA software environment. Beyond interoperable data exchange between the NCANDA and Synapse repository, this technology will support cross-repository queries and allow novel concept based queries using semantic annotation [36].

**Preliminary Results:** We designed and implemented the NCANDA Data Integration framework that fuses cognitive measures, clinical data, and neuroimaging data from five different collection sites in a version controlled and automated system [2]. We used Synapse [5], [6], an open source data workspace and repository that accommodates a wide variety of biomedical data and bioinformatics analyses, to distribute a repository of brain shape measures [39].

**Approach:** We will first extend the NCANDA software platform to support the NIDM data exchange specification developed in Aim 1.We will then create a Web Service API for the Cognitive Atlas ontology that will make automated semantic annotation of cognitive measures publicly accessible. Once these software packages are created and unit tested, we will conform the NCANDA data to the data exchange specifications using the Web Service API of the Cognitive Atlas ontology for metadata tagging. SRI will upload the NIDM compliant representation of the NCANDA data to the Synapse repository using the Synapse API. Finally, we confirm proper uploading of the primary data of NCANDA by applying the evaluation of Aim1 to the entire data set.

**Evaluation:** We will perform a set of queries to ensure that the proposed platform will enhance the query capabilities on the NCANDA data. Specifically, SRI will create a set of queries impossible to perform on the primary data such as “finding the dimensions of neuropsychological function related to substance use and their relation to brain shape measures”. Next Sage will apply those queries to the Synapse repository. We will view the evaluation, as a success should these queries return comprehensive findings across all subjects monitored by NCANDA.

**Risk and Alternatives**: Important neuroscience terms and concepts missing in the Cognitive Atlas ontology could jeopardize the query capabilities of the Synapse repository with respect to the NCANDA data. Should that be the case then we will include other existing ontologies into our framework, such as the PhenX Toolkit [40] and NDAR [3].

**C.4. Timeline and Milestones**
**Months 1-6:** a) Identify object models for neuroimaging and cognitive measures b) Draft object model specifications for neuroimaging measures c) draft object model specifications for cognitive mesures.
**Milestone 1:** Release a version of the Neuroimaging Data Model that supports a subset of the data and concepts used by NCANDA, Cognitive Atlas, and Synapse. Obtain the set of queries used for validations.
**Months 2-12:** a) Implement a software library to convert NCANDA data into the NIDM data exchange layer, b) Design and implement the Cognitive Atlas API c) Implement a data interoperability library for synapse to ingest the NIDM data exchange standard, and d) Extend the NCANDA data integration framework to support automated upload of the NCANDA imaging and cognitive measures
**Milestone 2:** Demonstrate the overall framework by issuing federated queries across multiple sources using the NIDM representation of each data source and uploading the NCANDA Baseline visit data to Synapse.

## References


[1] L. D. Johnston, P. M. OMalley, R. A. Miech, J. G. Bachman, and J. E. Schulenberg, “Monitoring the Future national survey results on drug use: 1975-2014: Overview, key findings on adolescent drug use. ,” 2015.

[2] T. Rohlfing, K. Cummins, T. Henthorn, W. Chu, and B. N. Nichols, “N-CANDA data integration: anatomy of an asynchronous infrastructure for multi-site, multi-instrument longitudinal data capture,” *Journal of the American Medical Informatics Association*, pp. amiajnl–2013–002367, 2013.

[3] D. Hall, M. F. Huerta, M. J. McAuliffe, and G. K. Farber, “Sharing Heterogeneous Data: The National Database for Autism Research,” *Neuroinformatics*, vol. 10, no. 4, pp. 331–339, 2012.

[4] R. A. Poldrack, A. Kittur, D. Kalar, E. Miller, C. Seppa, Y. Gil, D. S. Parker, F. W. Sabb, and R. M. Bilder, “The cognitive atlas: toward a knowledge foundation for cognitive neuroscience.,” *Front. Neuroinform.*, vol. 5, p. 17, 2011.

[5] J. M. J. Derry, L. M. Mangravite, C. Suver, M. D. Furia, D. Henderson, X. Schildwachter, B. Bot, J. Izant, S. K. Sieberts, M. R. Kellen, and S. H. Friend, “Developing predictive molecular maps of human disease through community-based modeling.,” *Nature genetics*, vol. 44, no. 2, pp. 127–130, Feb. 2012.

[6] L. Omberg, K. Ellrott, Y. Yuan, C. Kandoth, C. Wong, M. R. Kellen, S. H. Friend, J. Stuart, H. Liang, and A. A. Margolin, “Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas.,” *Nature genetics*, vol. 45, no. 10, pp. 1121–1126, Oct. 2013.

[7] A. R. Ferguson, J. L. Nielson, M. H. Cragin, A. E. Bandrowski, and M. E. Martone, “Big data from small data: data-sharing in the ‘long tail’ of neuroscience.,” *Nature neuroscience*, vol. 17, no. 11, pp. 1442–1447, Nov. 2014.

[8] K. Buckow, M. Quade, O. Rienhoff, and S. Y. Nussbeck, “Changing requirements and resulting needs for IT-infrastructure for longitudinal research in the neurosciences.,” *Neurosci. Res.*, Aug. 2014.

[9] D. Keator, J. Grethe, D. Marcus, B. Ozyurt, S. Gadde, S. Murphy, S. Pieper, D. Greve, R. Notestine, and H. Bockholt, “A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN),” *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 2, pp. 162–172, 2008.

[10] M. Mennes, B. B. Biswal, F. X. Castellanos, and M. P. Milham, “Making data sharing work: the FCP/INDI experience.,” *NeuroImage*, vol. 82, pp. 683–691, Nov. 2013.

[11] A. W. Toga, K. L. Crawford, Alzheimer’s Disease Neuroimaging Initiative, “The informatics core of the Alzheimer's Disease Neuroimaging Initiative.,” *Alzheimer“s & Dementia: The Journal of the Alzheimer”s Association*, vol. 6, no. 3, pp. 247–256, May 2010.

[12] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, “Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults.,” *J Cogn Neurosci*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007.

[13] I. B. Ozyurt, D. B. Keator, D. Wei, C. Fennema-Notestine, K. R. Pease, J. Bockholt, and J. S. Grethe, “Federated web-accessible clinical data management within an extensible neuroimaging database.,” *Neuroinformatics*, vol. 8, no. 4, pp. 231–249, Dec. 2010.

[14] R. C. Gur, J. Richard, P. Hughett, M. E. Calkins, L. Macy, W. B. Bilker, C. Brensinger, and R. E. Gur, “A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: Standardization and initial construct validation,” *Journal of neuroscience methods*, vol. 187, no. 2, pp. 254–262, Mar. 2010.

[15] P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde, “Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support.,” *J Biomed Inform*, vol. 42, no. 2, pp. 377–381, Apr. 2009.

[16] D. Marcus, T. Olsen, M. Ramaratnam, and R. Buckner, “The extensible neuroimaging archive toolkit,” *Neuroinformatics*, vol. 5, no. 1, pp. 11–33, 2007.

[17] J. B. Poline, J. L. Breeze, S. Ghosh, K. Gorgolewski, Y. O. Halchenko, M. Hanke, C. Haselgrove, K. G. Helmer, D. B. Keator, D. S. Marcus, R. A. Poldrack, Y. Schwartz, J. Ashburner, and D. N. Kennedy, “Data sharing in neuroimaging research.,” *Front. Neuroinform.*, vol. 6, pp. 9–9, 2012.

[18] S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.-A. Coleman, J. Copeland, S. Das, A. de Daruvar, P. de Matos, I. Dix, S. Edmunds, C. T. Evelo, M. J.

Forster, P. Gaudet, J. Gilbert, C. Goble, J. L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S. J. Ho Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C. E. Shamu, C. A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios, and W. Hide, “Toward interoperable bioscience data : Nature Genetics : Nature Publishing Group,” *Nature genetics*, vol. 44, no. 2, pp. 121–126, Feb. 2012.

[19] D. B. Keator, K. Helmer, J. Steffener, J. A. Turner, T. G. Van Erp, S. Gadde, N. Ashish, G. A. Burns, and B. N. Nichols, “Towards structured sharing of raw and derived neuroimaging data across existing resources,” *NeuroImage*, vol. 82, pp. 647–661, 2013.

[20] F. Belleau, F. Belleau, N. Tourigny, N. Tourigny, B. Good, B. Good, J. Morissette, and J. Morissette, “Bio2RDF: A Semantic Web atlas of post genomic knowledge about human and mouse,” *Data Integration in the Life …*, 2008.

[21] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L Whitwell, C. Ward, A. M. Dale, J. P. Felmlee, J. L. Gunter, D. L. G. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. S. DeCarli, G. Krueger, H. A. Ward, G. J. Metzger, K. T. Scott, R. Mallozzi, D. Blezek, J. Levy, J. P. Debbins, A. S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, and M. W. Weiner, “The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods.,” *Journal of magnetic resonance imaging: JMRI*, vol. 27, no. 4, pp. 685–691, Apr. 2008.

[22] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, S. Della Penna, D. Feinberg, M. F. Glasser, N. Harel, A. C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. E. Petersen, F. Prior, B. L. Schlaggar, S. M. Smith, A. Z. Snyder, J. Xu, E. Yacoub, WU-Minn HCP Consortium, “The Human Connectome Project: a data acquisition perspective.,” *NeuroImage*, vol. 62, no. 4, pp. 2222–2231, Oct. 2012.

[23] R. A. Poldrack, D. M. Barch, J. P. Mitchell, T. D. Wager, A. D. Wagner, J. T. Devlin, C. Cumba, O. Koyejo, and M. P. Milham, “Toward open sharing of task-based fMRI data: the OpenfMRI project.,” *Front. Neuroinform.*, vol. 7, p. 12, 2013.

[24] O. Brazhnik, O. Brazhnik, J. F. Jones, and J. F. Jones, “Anatomy of data integration,” *J Biomed Inform*, vol. 40, no. 3, pp. 252–269, 2007.

[25] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Scientific American*, vol. 284, no. May, pp. 1–4, May 2001.

[26] D. Botstein, J. M. Cherry, M. Ashburner, C. A. Ball, J. A. Blake, H. Butler, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene Ontology: tool for the unification of biology - Nature Genetics,” *Nature genetics*, vol. 25, no. 1, pp. 25–29, May 2000.

[27] C. Rosse and J. L. V. Mejino, “A reference ontology for biomedical informatics: the Foundational Model of Anatomy.,” *J Biomed Inform*, vol. 36, no. 6, pp. 478–500, Dec. 2003.

[28] P. Karp, “Pathway databases: a case study in computational symbolic theories,” *Science*, vol. 293, no. 5537, p. 2040, 2001.

[29] M. Galdzicki, K. P. Clancy, E. Oberortner, M. Pocock, J. Y. Quinn, C. A. Rodriguez, N. Roehner, M. L. Wilson, L. Adam, J. C. Anderson, B. A. Bartley, J. Beal, D. Chandran, J. Chen, D. Densmore, D. Endy, R. Grünberg, J. Hallinan, N. J. Hillson, J. D. Johnson, A. Kuchinsky, M. Lux, G. Misirli, J. Peccoud, H. A. Plahar, E. Sirin, G.-B. Stan, A. Villalobos, A. Wipat, J. H. Gennari, C. J. Myers, and H. M. Sauro, “The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology.,” *Nature Biotechnology*, vol. 32, no. 6, pp. 545–550, Jun. 2014.

[30] S. D. Larson and M. E. Martone, “NeuroLex.org: an online framework for neuroscience knowledge.,” *Front. Neuroinform.*, vol. 7, p. 18, 2013.

[31] D. Gardner, H. Akil, G. Ascoli, D. Bowden, W. Bug, D. Donohue, D. Goldberg, B. Grafstein, J. Grethe, and A. Gupta, “The neuroscience information framework: a data and knowledge environment for neuroscience,” *Neuroinformatics*, vol. 6, no. 3, pp. 149–160, 2008.

[32] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. L. Chepelev, J. Cruz-Toledo, R. Nicholas, D. Rio, G. Duck, and L. I. Furlong, “The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery.,” *J. Biomedical Semantics*, vol. 5, p. 14, 2014.

[33] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, and C. J. Mungall, “The OBO Foundry: coordinated evolution of ontologies to support

biomedical data integration," *Nature Biotechnology*, vol. 25, no. 11, pp. 1251–1255, 2007.
[34] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems," *J Biomed Inform*, vol. 41, no. 5, Oct. 2008.
[35] B. N. Nichols, J. L. Mejino Jr, L. Detwiler, T. T. Nilsen, M. E. Martone, J. A. Turner, D. L. Rubin, and J. F. Brinkley, "Neuroanatomical domain of the foundational model of anatomy ontology.," *J. Biomedical Semantics*, vol. 5, p. 1, 2014.
[36] J. A. Turner, J. L. V. Mejino, J. F. Brinkley, L. T. Detwiler, H. J. Lee, M. E. Martone, and D. L. Rubin, "Application of neuroanatomical ontologies for neuroimaging data annotation.," *Front. Neuroinform.*, vol. 4, 2010.
[37] D. L. McGuinness and F. Van Harmelen, "OWL web ontology language overview," *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004.
[38] F. Manola and E. Miller, Eds., "RDF Primer," *www.w3.org*, 10-Feb-2004. [Online]. Available: http://www.w3.org/TR/rdf-primer/. [Accessed: 13-Dec-2013].
[39] A. Klein and J. Tourville, "101 labeled brain images and a consistent human cortical labeling protocol.," *Frontiers in Neuroscience*, vol. 6, p. 171, 2012.
[40] P. J. Stover, W. R. Harlan, J. A. Hammond, T. Hendershot, and C. M. Hamilton, "PhenX: a toolkit for interdisciplinary genetics research," *Current Opinion in Lipidology*, vol. 21, no. 2, pp. 136–140, Apr. 2010.