Ontology Driven Data Integration for Autism Research

Lynn Young[1], Samson W. Tu[2], Lakshika Tennakoon[2], David Vismer[1], Vadim Astakhov[3], Amarnath Gupta[3], Jeffrey S. Grethe[3], Maryann E. Martone[3], Amar K. Das[2], and Matthew J. McAuliffe[1]

[1]Division of Computational Bioscience, CIT, U.S. National Institutes of Health, Bethesda, MD 20892, [2]Center for Biomedical Informatics Research, Stanford University, Stanford CA 94305-5479, and [3]Center for Research in Biological Systems, University of California, San Diego, San Diego, CA 92093, {lynny, mcmatt, vismerd}@mail.nih.gov, {swt, lakshika, das}@stanford.edu, {astakhov, jgrethe, maryann}@ncmir.ucsd.edu, and gupta@sdsc.edu

Abstract

Autism Spectrum Disorder is an inherently complex phenomenon requiring large studies of many different types to further understanding of its causes. The National Database for Autism Research (NDAR) is being constructed to aid in this effort by providing a means for researchers to share and integrate data. An autism ontology drafted by a group at Stanford is being incorporated for use by NDAR to allow semantic data integration. The architecture upon which NDAR is built - the UCSD Developed Data Integration Environment - supports the use of this autism ontology, including annotation of data with ontological concepts and ontology enhanced queries on databases, both central and federated.

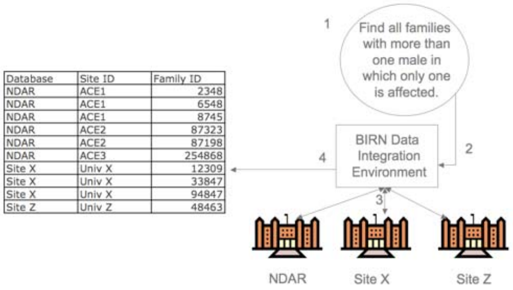
1. Introduction

Autism, first described by Leo Kanner in 1943 [1], comprises problems in social interactions, difficulties in communication, and repetitive behavior. Currently, autism is seen as a range of disorders, known as Autism Spectrum Disorder or ASD. The spectrum includes Asperger syndrome and Pervasive Developmental Disorder - Not Otherwise Specified [2]. Estimates of prevalence depend on diagnostic criteria, age of population included in the study, and whether the study sample is located in a rural or metropolitan area. An estimate based on a pool of studies showed that for typical autism 7.1 per 10,000 individuals are affected, while for ASD, 20.0 per 10,000 are affected [3].

Research in autism is currently focused in the areas of cognition, clinical phenotype, treatment, social function, brain imaging, and genetics, to name a few [4]. A large portion of autism research funding goes to assessing individuals for ASD. These assessments can then be used in areas such as genomics and functional neuroimaging to search for correlations between data from these experiments and ASD individuals. In recent years, studies collecting both imaging and genomics data have begun to appear. However, due to the heterogeneous nature of the disorder, the number of individuals in the study must be larger than that in other clinical studies.

Useful for a researcher would be access to studies from other labs such that he or she could begin to understand what type of process would be required to combine data from multiple studies, thus increasing the study population. To facilitate integration of data from experiments across various areas of autism research and to allow researchers to combine subjects across institutions, the National Database of Autism Research (NDAR) was created (http://ndar.nih.gov). The system allows researchers to submit autism data both for sharing and for integration with other studies. Data integration is supported either by direct submission of data to the NDAR Central Repository or by data federation. Data federation allows a site, wishing to maintain data in its own repository, to expose the sharable portion of its information to NDAR users in such a way that the user need not know how to access the additional sites (Figure 1). For NDAR, data federation is accomplished using the UCSD developed Data Integration Environment which was originally part of the Biomedical Informatics Research Network (BIRN)[5,6,7,8]. NDAR was built on the original BIRN.

978-1-4244-4878-4/09/$25.00 ©2009 IEEE



Database	Site ID	Family ID
NDAR	ACE1	2348
NDAR	ACE1	6548
NDAR	ACE1	8745
NDAR	ACE2	87323
NDAR	ACE2	87198
NDAR	ACE3	254868
Site X	Univ X	12309
Site X	Univ X	33847
Site X	Univ X	94847
Site Z	Univ Z	48463

1
Find all families with more than one male in which only one is affected.
2
BIRN Data Integration Environment
3
4
NDAR
Site X
Site Z

Figure 1. Data Integration and Federation
Federation is accomplished via the UCSD Data Integration Environment, where each site registers a data source.

The UCSD Data Integration Environment uses a grid computing architecture to support collaborative research. It supports distributed databases and file systems and provides user interfaces, application programming interfaces, and web services with the goal of making data federation transparent to researchers. The data integration environment (Figure 2) comprises ontology for semantic integration; a mapping of the ontology to the federated data; a means to expose data sources to the grid; and middleware to mange federated queries and data extraction. Although the focus of the ontology, initially, has been on neuroimaging, a process exists for importing new ontologies such as ontology for autism [9].

The Das group at Stanford Biomedical Informatics Research (BMIR) has developed an approach for using ontologies and data models for querying and reasoning about phenotypes, using autism as an example [10]. Many ontologies are created without access to data, such that when a researcher wants to apply the ontology for enriching query results, he or she may be faced with the daunting task of figuring out how the data maps to the ontology. The approach of this group is to include the data model as a part of the design process such that access to data via the ontology becomes transparent. Additionally, their use of semantic web technologies to encode the ontology enables reasoning to be performed computationally.

A major benefit to autism research of these components working together is the vision of a phenotype catalog [11] to not only index and store phenotypes used in autism research, but also to use them to query autism data to generate lists of subjects satisfying the phenotype, as well as what type of measurement data are available for these subjects. Additionally, a long term goal would be the implementation of algorithmically determined phenotypes such as those determined by principal components analysis of autism diagnostic data [12, 13].

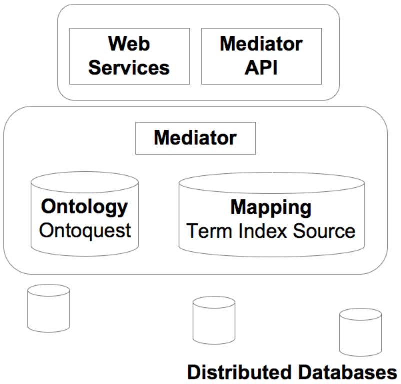


Figure 2. UCSD Data Integration Environment
The user query interface is built on the UCSD Data Integration Environment. Data from distributed databases can be mapped to the ontology such that the mapping is stored in the environment. Web services provide a connection between the user interface and the ontology and mapping information. The Mediator provides access to the data.

2. Autism Ontology

At a high level the autism ontology [10] integrates (1) phenotypic concepts in autism ; (2) an information model that represents research and clinical data; and (3) abstraction rules that relate observable data with phenotypic concepts. The phenotypic concepts have been initially chosen from the literature and from the DSM-IV autism definition. The description can be thought of in terms of object oriented design, in which each phenotype is a class or subclass in a disease or disposition hierarchy.

The data model NDAR uses to describe ASD is based on diagnostic instruments. At this level of detail, the ontology models the assessment results of instruments as subclasses of some information content entity and the instrument items as properties of these subclasses using the Web Ontology Language (OWL). The beauty of this design is that complex phenotypes defined as combinations of scores over more than one instrument or algorithmic operations on scores and sub scores can be represented in the ontology as abstraction rules. The rules and queries based on them can be encoded in the ontology using the Semantic Web Rule Language (SWRL) and the Semantic Query-enhanced Web Rule Language (SQWRL) to leverage the ontology and its reasoning capability to extract data from NDAR [14].

3. UCSD Developed Data Integration Environment

Suppose one were to perform a keyword search for information, and suppose that in addition to getting a search result with information containing the keyword, one received related terms provided by the search software. And suppose that the software had the ability to reason using ontology rules to combine the terms to create a search useful to the user but that did not initially occur to him or her. And suppose the information could be obtained from sources of which the user was not even aware. Such a solution is provided (and under development) in the UCSD Data Integration Environment. The core of this environment comprises components for 1) exposing data for federation; 2) mapping data to ontologies; and 3) querying and accessing the data.

The ontology for this environment is BIRNLex/NeuroLex. With the funding of the Neuroscience Information Framework (NIF) project, led by UCSD, all work on the BIRN ontologies will be subsumed by the NIF. The BIRN ontologies form the core set of vocabulary resources (NeuroLex) of the NIF. NeuroLex is the outward facing version of the ontology to which the neuroscience community can provide input via a wiki (http://neurolex.org). The core ontology is the NIF standardized ontology (NIFSTD) which is maintained in OWL. It is included in the National Center for Biomedical Ontology (NCBO) BioPortal (http://bioportal.bioontology.org). In keeping with the Open Biological Ontology (OBO) community best practices, NIFSTD was built using a multitude of resources including the Unified Medical Language System (UMLS) [9]. An example of an ontology founded on NIFSTD is Disease Ontology (http://openccdb.org/wiki/index.php /Disease_Ontology).

To use the UCSD Data Integration Environment, each site registers data sources with the data integration framework. An important component of this framework is the mediator - a component that decomposes a user query into subqueries, sends it to different relational data sources, and assembles partial results into complete results that are returned to the users. This component wraps the data source such that a common API can be used to access all sources. Tables, fields, and values can then be mapped to the ontology. The mappings are stored in the Term Index Source. Web services allow the implementation of user interfaces to this environment. A web service method can look up the keyword in Ontoquest, an ontology management module that allows a user to access and perform graph queries on it. A web service method can use the resulting concept identifiers to look up ontological relationships between concepts to expand the search. An additional method looks up the resulting data mapped to these concepts. Currently, the data federation API allows some additional logic to evaluate queries semantically, but concerns exist regarding the amount of data for transfer for this step. For example, a phenotype may involve the average of several scores from a diagnostic instrument. If this function is performed by the federation engine, it would need access to all of the data to calculate the average before the selection of those individuals satisfying the phenotype rule could be returned to the user. For this reason, initially, additional semantic query evaluation applications may be mirrored at each source. An example for NDAR is shown in Figure 3.

4. Application to NDAR

NDAR will provide users the ability to generate a list of subjects satisfying a particular phenotype. Long term plans also include ontology enhanced queries to identify available data. To gain access to the autism ontology through the UCSD Data Integration Environment, NDAR requested that the autism ontology be imported into the BIRNLex/NeuroLex. The curation team at the UCSD NIF will assign the concept identifiers during this process. Since the team can assign concept identifiers to subclass properties, the initial mapping of data to concept identifiers will be straightforward (see section “Autism Ontology”). By making all of the autism ontology classes available through the NeuroLex, they can be incorporated into other neuroscience resources without the need for extensive cross-mapping.

Initially, for the phenotype catalog, the reasoning step, necessary to generate the phenotypes, will be applied at each data source. Reasoning will be implemented in SQL and the list of subjects satisfying the criteria will be stored in a database view. In the intermediate term, reasoning will be applied using Dynamic DataMaster, the API from O’Connor, et al. [14] that allows opportunistic loading of database data into the OWL/SWRL environment. The database schema will be extended such that a phenotype catalog table can be connected to a table of subjects via a join table. A long term goal is to centralize reasoning by executing rules in the UCSD data integration environment, if the size of data transfer is manageable.

Access to the data will be provided using the NDAR Query Tool, where a user can select a phenotype of interest. For example, clicking on a hyperlink to an phenotype could return a list of subjects satisfying the phenotype and including links to available data. Only those subjects the user has permission to access would appear in the list.

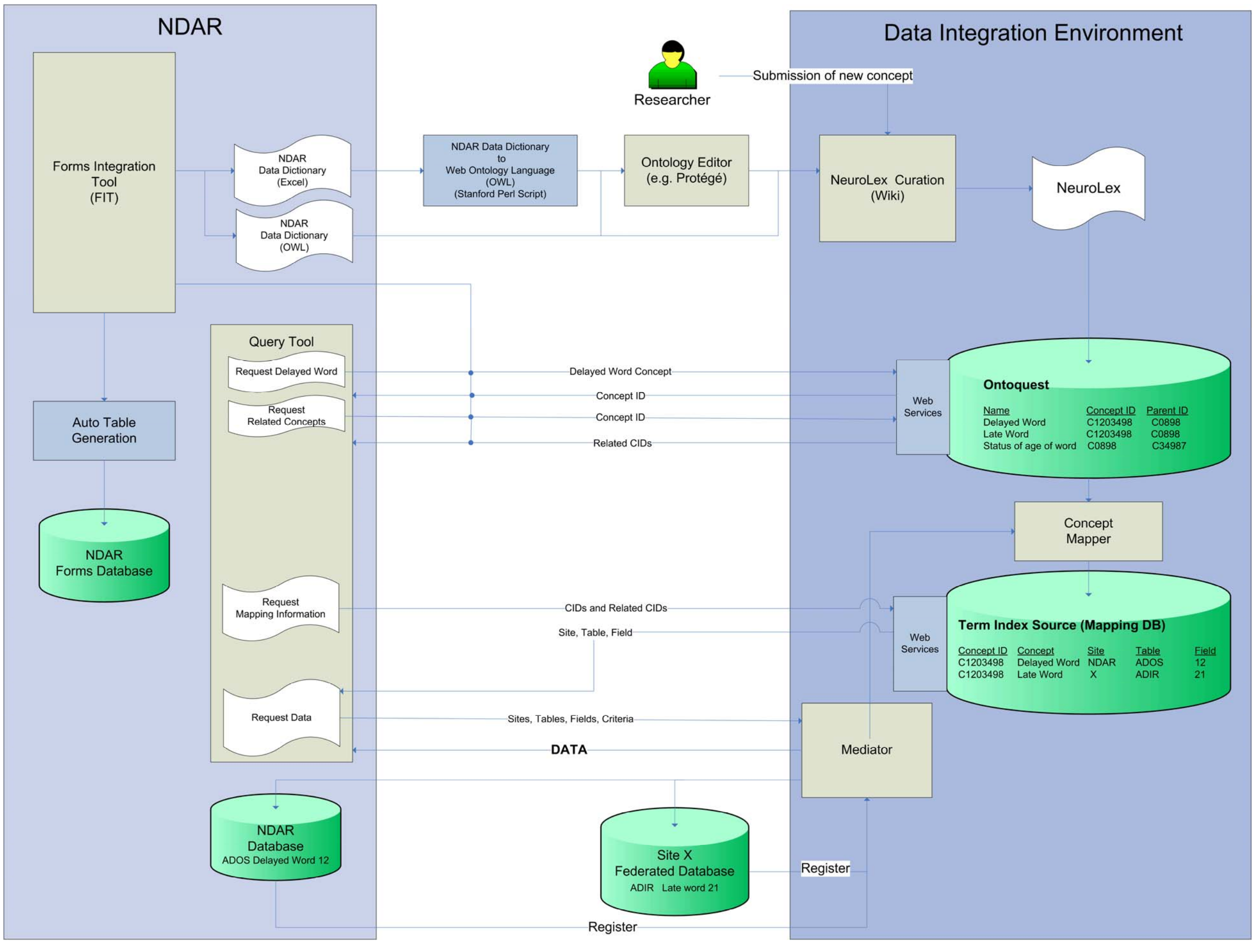

NDAR
Data Integration Environment
Researcher
Submission of new concept
Forms Integration Tool (FIT)
NDAR Data Dictionary (Excel)
NDAR Data Dictionary (OWL)
NDAR Data Dictionary to Web Ontology Language (OWL) (Stanford Perl Script)
Ontology Editor (e.g. Protégé)
NeuroLex Curation (Wiki)
NeuroLex
Query Tool
Request Delayed Word
Request Related Concepts
Delayed Word Concept
Concept ID
Concept ID
Related CIDs
Web Services
Ontoquest
Name Concept ID Parent ID
Delayed Word C1203498 C0898
Late Word C1203498 C0898
Status of age of word C0898 C34987
Auto Table Generation
NDAR Forms Database
Concept Mapper
Request Mapping Information
CIDs and Related CIDs
Site, Table, Field
Web Services
Term Index Source (Mapping DB)
Concept ID Concept Site Table Field
C1203498 Delayed Word NDAR ADOS 12
C1203498 Late Word X ADIR 21
Request Data
Sites, Tables, Fields, Criteria
DATA
Mediator
NDAR Database
ADOS Delayed Word 12
Site X Federated Database
ADIR Late word 21
Register
Register

Figure 3. Example Ontology Enhanced Query Environment for NDAR
NDAR could build a Forms Integration Tool to let researchers represent new diagnostic instruments in the system. This tool would store the information in an NDAR database and also generate an entry for the autism ontology. This entry would be submitted to the NeuroLex curation team for inclusion in the NeruoLex. Alternatively, autism researchers could submit concepts directly to NeuroLex. The NeuroLex is transformed to a relational database structure (Ontoquest) for programmatic access via web services. Two data sources are shown at the bottom of the figure. These must be registered with the mediator for programmatic access. At this point the Concept Mapper displays both the ontology concepts and the registered data sources, allowing the user to map table names, field names and field values to the ontology. This mapping information is stored in the Term Index Source. The NDAR Query Tool would then be able to request a lookup of the concept "delayed word", for example. The web service would return the concept identifier from Ontoquest. Next the query tool could expand the list of concepts by using a web service to Ontoquest to find related terms in the hierarchy. This list of concepts could be given to a web service to the Term Index Source to discover any mapping information to data. The names of data source sites, tables, and fields would be returned, allowing the query tool to display the relevant data. Similarly, the Forms Integration Tool could access the Ontoquest web services to annotate new forms with concepts from the NeuroLex.

4.1. Proof of Concept

The focus of the autism ontology is phenotype. One section describes phenotype level such as measures for language acquisition (Figure 4). One of the measures is the age when words are spoken. Three categories are given for the status: delayed word, no word, or non delayed word [15]. As an example, we can use the phenotype level called "delayed word" which has the following SWRL rule [10]:

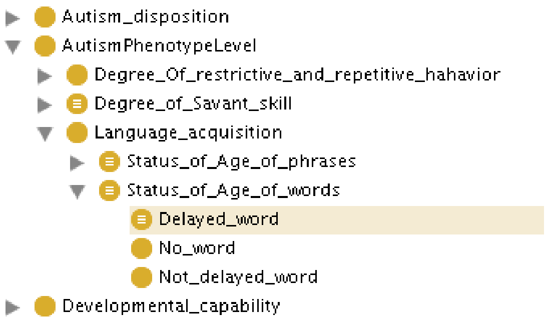
ADI_2003_result(?assessment) &
acqorlossoflang_aword(?assessment,?wordage) &
swrlb:greaterThan(?wordage, 24) &
subject_id(?assessment, ?subjectId) &
orgtax:Human(?subject) &
subject_id(?subject, ?subjectId) →
birn_obo_ubo:bearer_of(?subject, Delayed_word)

We could create a corresponding database view with each field corresponding to a predicate name in an atom in the rule. For this example, we use the same order as the atoms in the rule. The field names are

Instrument
Age of first single words (if ever used)
Flag for age greaterThan 24
Global Unique Identifier (NDAR GUID)
Organism
Flag for GUID matching human
Phenotype.

The creation of the view would use constraints on the variable bindings. The view would be registered with the mediator such that it could be accessed by NDAR users with the NDAR Query Tool. This means that the rule would be executed when the user selects this view for a query result.

Alternatively, the database schema could be extended to accommodate a table of phenotype records, and an additional table could be added to store a list of subjects and the phenotypes which each bears. Such a table could be updated on a daily basis at each source. The advantage of this approach is that the queries would return the data much faster, especially if multiple phenotypes were queried; however, subjects added between updates would not be in the query results until the next day. Eventually, the ideal would be to apply the rules centrally in the UCSD data integration environment.



- ▶ Autism_disposition
- ▼ AutismPhenotypeLevel
  - ▶ Degree_Of_restrictive_and_repetitive_hahavior
  - ▶ Degree_of_Savant_skill
  - ▼ Language_acquisition
    - ▶ Status_of_Age_of_phrases
    - ▼ Status_of_Age_of_words
      - Delayed_word
      - No_word
      - Not_delayed_word
- ▶ Developmental_capability

Figure 4. The context of the "Delayed word" phenotype level is shown.

For expediency in implementing this proof of concept, the rules are translated into SQL, but some of the more complex rules will involve recursive relationships which are not handled well by SQL. Thus, generation of phenotype in future releases of NDAR will turn to more efficient solutions in implementing rule execution.

For the federation step, in this example, we would submit the term "delayed word" for curation in the NeuroLex, including its definition and information about where it fits into the NeuroLex. The curators would assign a NeuroLex Concept ID to the term "delayed word" and list it under the module "NIF Phenotype". In this way, if site X has the term "delayed word", its subject matter expert can map it to the same NeuroLex Concept ID, and results of views from both NDAR and site X will be returned to a user looking for subjects satisfying the endophenotype "delayed word". Additionally, if Site X had the term "late word" which was determined to be the same concept, the expert could map this term to the same concept identifier, so that data for this concept could be accessed. Note also that by removing the class explicitly from the Autism Ontology, it becomes available to any other disorder that results in "delayed word".

The Ontoquest system is a database system specifically designed to serve information contained in OWL ontologies efficiently so that some of the views that can be generated from reasoning on the ontologies can be materialized without having to do it on the fly. The current version returns parents, children, synonyms and other relationships attached to classes such as "part of". Future versions of Ontoquest will also incorporate the ability to utilize rules such as that outlined above, to offer that as an option to someone who is looking for "delayed word", for example. Because the ultimate goal of the NIF and projects like NDAR is to allow users to search across information sources, regardless of origin or perceived relevance, we envision a system whereby all communities compose composite entities such as "delayed word" and provide formal definitions for their particular community as shown above, for example, referencing the specific test instruments and scores that constitutes delayed language in their field. When one searches for "delayed word", these specific views would be offered to the user.

In this example, the phenotype term corresponds to the view. After exposing this view to the data integration environment, by registering it with the mediator, the Concept Mapper could be used to link the view (and thus our data) to the "delayed word" concept in the ontology. The Autumn 2008 release of NDAR contains the implementation of this example phenotype view that the user can select. The result of selecting this view is the list of subjects satisfying the phenotype and a display of the data for the other fields in the view. This is the beginning of an NDAR phenotype catalog for autism research.

5. Discussion

The most common source for phenotypes used in autism is the research literature. Often the details of the phenotype may be somewhat vague. The SWRL component of the ontology would provide a standardized language for encoding phenotype. NDAR could provide an application to allow researchers to define the phenotype being used - to define it in terms of the elements in our data dictionary which reflects the assessments and their items. This application could then convert the phenotype to SWRL. In some cases, the rules would be complex algorithms, such as principal components analyses, that could be implemented in NDAR and run periodically.

NDAR could also assign accession numbers to phenotype and associate literature citations with them. In this way, a researcher could simply type a citation into NDAR to retrieve the phenotypes therein and then get the pertinent list of subjects. Alternatively, researchers could get phenotype accession numbers from NDAR prior to publication (keeping them private until that time) such that the phenotype accession number could be included in the publication itself. The reader could simply search NDAR using the accession number to retrieve the phenotype and associated data.

6. Conclusion

Since autism is defined as a spectrum disorder, autism phenotype is necessarily complex. Provision of a standard representation of autism phenotypes should prove beneficial to researchers in this field. In turn, standard phenotype representations could drive data integration. By increasing the number of subjects in studies, correlations of genomics and imaging data with clinical assessments data may also increase. Ultimately, clinical genetics or imaging tests could then be run to detect susceptibility to autism.

Acknowledgements

William Bug (1961-2008) was instrumental in this project. This work was supported by 1) the Intramural Research Program of the NIH, Center for Information Technology (LY, DV, MJM), 2) NLM 1P41LM007885 (SWT, LT, AKD), and 3) NINDS RO1NS058296, NCRR RR04050, and RR08605 (VA, AG, JSG, and MEM).

10. References

[1] L. Eisenberg, "Images in Psychiatry", *Am. J. Psychiatry* 151, 1994, p 751.

[2] S.H.N. Willemsen-Swinkels, and J.K. Buitelaar, "The autistic spectrum: subgroups, boundaries, and treatment", *Psychiatr Clin N Am* 25, 2002, pp. 811–836.

[3] J.G. Williams, J.P. Higgins, and C.E. Brayne, "Systematic review of prevalence studies of autism spectrum disorders", *Arch Dis Child* 91, 2006, pp. 2-5.

[4] IMFAR 2007 Abstracts website: http://www.autism-insar.org/docs/IMFAR2007_Program.pdf

[5] M.E. Martone, A. Gupta, and M.H. Ellisman, "E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains", *Nat Neurosci*, 7, 2004, pp. 467-472.

[6] B. Ludäscher, A. Gupta, and M.E. Martone, "Model-Based Information Integration in a Neuroscience Mediator System", *Proceedings of the 26th VLDB*, 2000, pp. 639-642.

[7] A. Gupta, B. Ludäscher, and M.E. Martone, "An Extensible Model-Based Mediator System with Domain Maps", 17th Intl. Conf. on Data Engineering (ICDE), 2001.

[8] V. Astakhov, A. Gupta, J. Grethe, E. Ross, D.R. Little, A. Yilmaz, M.E. Martone, X. Qian, S. Santini, and M.H Ellisman, "Semantically Based Data Integration Environment for Biomedical Research", *19th IEEE International Symposium on CBMS* 2006, 2006, pp. 171-176.

[9] W.J. Bug, G.A. Ascoli, J.S. Grethe, A. Gupta, C. Fennema-Notestine, A.R. Laird, S.D. Larson, D. Rubin, G.M. Shepherd, J.A. Turner, and M.E. Martone, "The NIFSTD and BIRNLex vocabularies: Building Comprehensive Ontologies for Neuroscience", *Neuroinformatics* 6, 2008, pp. 175-194.

[10] S.W. Tu, L. Tennakoon, M. O'Connor, R. Shankar, and A. Das, "Using an Integrated Ontology and Information Model for Querying and Reasoning about Phenotypes: The Case of Autism", *AMIA Annu Symp Proc.* 2008, pp.727–731.

[11] NIH Grants website: http://grants.nih.gov/grants/guide/rfa-files/RFA-MH-07-080.html

[12] C. Lord, B.L. Leventhal, and E.H. Cook Jr., Quantifying the phenotype in autism spectrum disorders", *Am J of Med Gen Part B.* 105, 2001, pp. 36–38.

[13] O. Tadevosyan-Leyfer, M. Dowd, R. Mankoski, B. Winklosky, S. Putnam, L. McGrath, H. Tager-Flusberg, and S.E. Folstein, "A Principal Components Analysis of the Autism Diagnostic Interview-Revised", *J Am Acad Child & Adolescent Psych* 42, 2003, pp. 864-872.

[14] M.J. O'Connor, R. Shankar, C. Nyulas, S. Tu, A. Das, "Developing a Web-Based Application using OWL and SWRL", AAAI Spring 2008 Symposia website: http://www.aaai.org/Papers/Symposia/Spring/2008/SS-08-01/SS08-01-012.pdf

[15] V. Hus, A. Pickles, E. Cook Jr., S. Risi, C. Lord, "Using the Autism Diagnostic Interview—Revised to Increase Phenotypic Homogeneity in Genetic Studies of Autism", *Biological Psychiatry*, 61, 2007, pp. 438-448.