

Designing an Ontology for the ENIGMA Neuroscience Collaboration

MiHyun Jang

Information Sciences Institute
University of Southern California

Contact: amyj08.2000@gmail.com

December 15, 2017

USC/ISI Technical Report

ISI-TR-723

ABSTRACT

The ENIGMA neuroscience collaboration has numerous working groups, each with many members that contribute diverse datasets. The lack of a data capture system makes it difficult for ENIGMA leadership to keep track of the various participants, investigators, leads, and members of working groups, projects, and cohorts. The absence of a system makes it challenging to keep track of the data used by each working group. To solve this problem, the ENIGMA Organic Data Science (ENIGMA-ODS) framework is being created to manage information about the collaborative aspects of ENIGMA. ENIGMA-ODS is supported by an ontology that establishes standard properties for various concepts within the ENIGMA network while allowing ENIGMA members to add additional properties as they see fit.

ISI-TR-723

Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey CA 90292, USA

1-310-822-1511

1. INTRODUCTION

ENIGMA is a global network of biomedical researchers in imaging genomics, neuroscience, and psychiatry [Thompson et al 2014]. Headquartered in University of Southern California, ENIGMA includes many researchers around the world. Within ENIGMA, there are many ongoing projects with a lot of diversity in their data in terms of the genomic and brain scan data types that are being used and the methods (protocols) to collect them. The ENIGMA Organic Data Science (ENIGMA-ODS) framework is a platform to manage information about the collaborative aspects of ENIGMA [Jang et al 2017]. ENIGMA-ODS reuses the LinkedEarth platform developed previously for collaborative paleoclimate research [Gil et al. 2017], developed as an extension of Semantic MediaWiki with a custom interface for end users to enter metadata.

This report describes an ontology for the ENIGMA-ODS framework to describe the concepts and data as well as all the properties of each concept necessary to support the ENIGMA collaboration. Since ENIGMA is very complex, a level of detail that explained most of the properties but took a reasonable time to map out had to be accomplished. Once the ontology is integrated onto the ENIGMA wiki, ENIGMA members can input information and metadata. The ontology allows for easy queries that can be answered by calling up various properties as well as the organization of information. Because the ENIGMA wiki is easy to use, members of ENIGMA who do not have any prior knowledge or experience with semantic representations will be able to populate the ontology easily.

The ontology was created using Protégé, an open-source ontology editor [Musen 2015; Protégé 2017]. The ontology represents core concepts and properties in ENIGMA. It can be extended in the future as additional requirements are collected.

2. PROBLEM ADDRESSED

Currently, information about the various projects in ENIGMA is recorded by each working group separately with no uniform system. As a result, researchers are often unaware of what other working groups are doing. In addition, the lack of an information system prevents ENIGMA researchers from easily querying the information, which would allow principal investigators and people of other roles to find out if other projects have the data that they seek. The members of ENIGMA need to find information that is currently not uniformly organized.

Figure 1 gives an overview of the complexity of the ENIGMA collaboration. One person can have a certain role in a particular working group, be the senior lead of 3 projects, a junior lead of another, and contribute to a fifth project in a minor role. The same person can also be the principal investigator (PI) of one or multiple cohorts, which are datasets collected from a target population (e.g. data collected from a population of schizophrenia patients following a carefully designed acquisition protocol). A cohort can be part of multiple working groups and projects, a

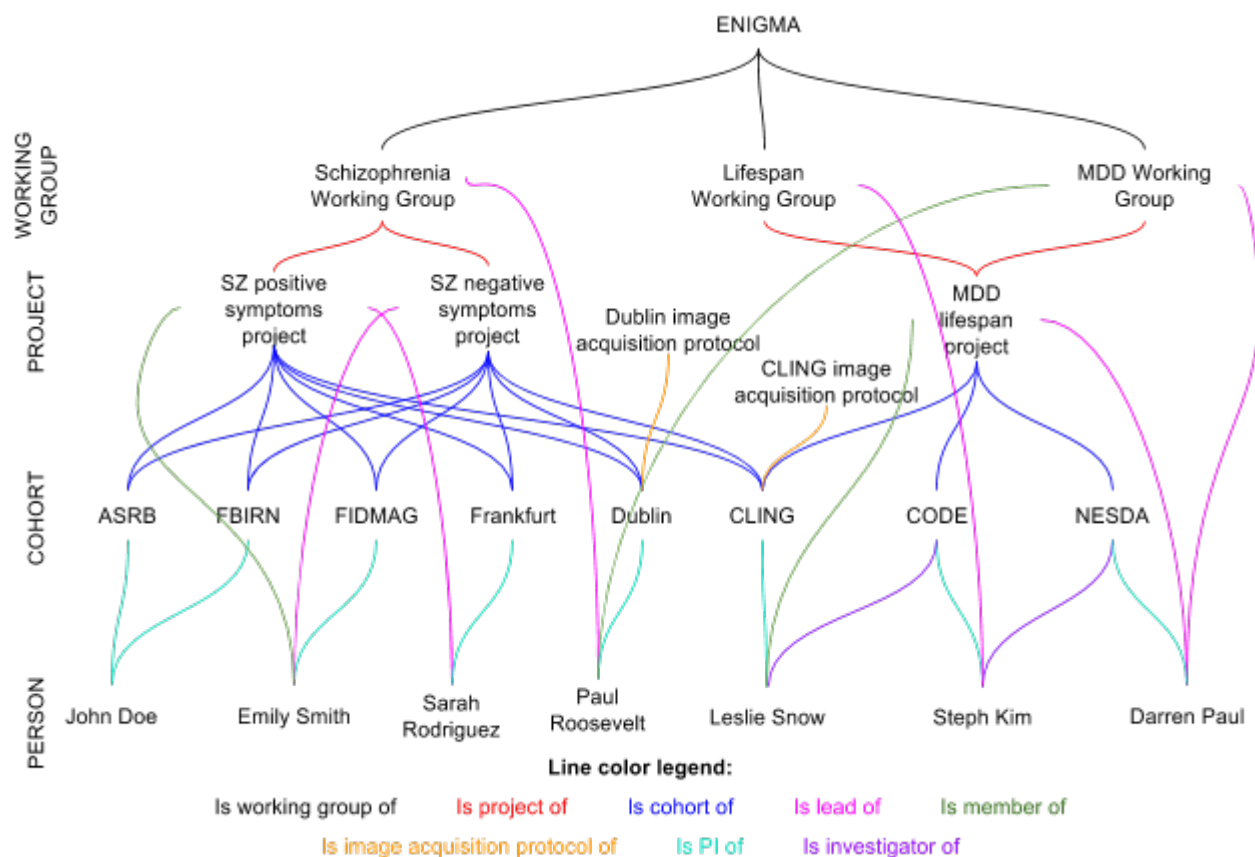


Figure 1: An illustration of the complexity of the ENIGMA collaboration. The names are not actual ENIGMA members, only examples.

project can be part of multiple working groups and have collected data from many cohorts, and a working group can have many projects and cohorts that may or may not be related to other working groups and projects. All this information can also change over time, as people move to new institutions or new projects are formulated. Working groups last a long time, while projects last a finite amount of time and typically result in a publication. Since spreadsheets and emails are used to record and collect information, the information can get quite unmanageable. The sheer amount of information within ENIGMA makes it difficult for members to keep track of the data collected and the relationships between people, working groups, projects, and cohorts.

Our goal is to develop a semantic representation of information about ENIGMA that is currently not captured. Researchers in projects and working groups will have the ability to easily query and aggregate information. A semantic representation will also allow for people to easily browse information about various aspects of ENIGMA, such as people's participation in projects and data collection procedures. An information system that incorporates semantic principles will help organize the data and help people in certain positions easily manage and search through the complicated ENIGMA collaboration.

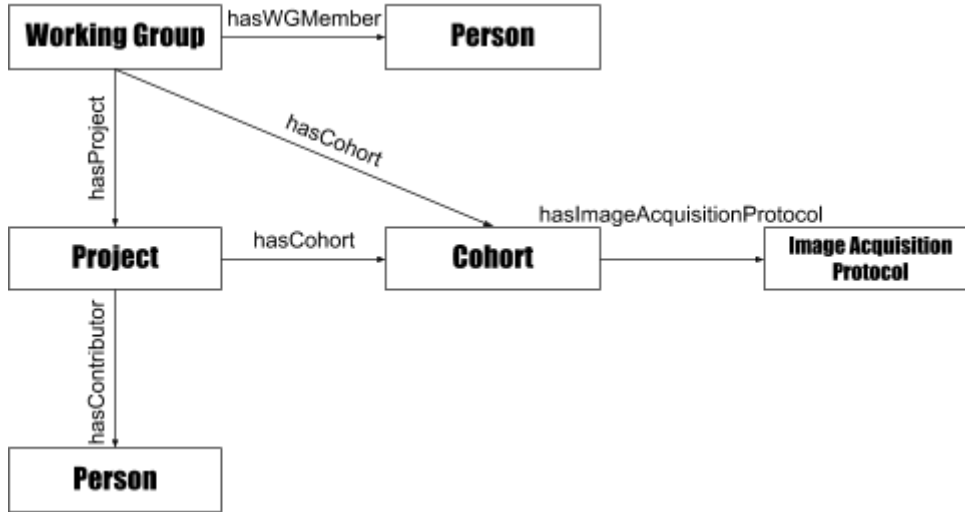


Figure 2: Core concepts in ENIGMA and relationships among them.

2.1 Requirements

The main requirements for the ontology were twofold: 1) it had to correctly represent all the different concepts within ENIGMA and their main properties and 2) it had to correctly describe the authors of ENIGMA project papers and the information of the main data tables in supplementary materials of those papers. The main concepts of ENIGMA are as follows: Working Group, Project, Cohort, Image Acquisition Protocol, and Person. Figure 2 gives a visual representation of these concepts and the relationships among them. The authors and information from the data tables can be described by using these concepts and relationships of ENIGMA.

There are over 30 working groups within ENIGMA that focus on various functions, diseases, and structures of the brain. The *working groups* analyze data about cohorts. Examples of working groups include the Schizophrenia (SZ) Working Group, the Genome-wide Association Study (GWAS) Working Group, and the Lateralization Working Group. Each working group has many members who take on various roles, such as *leads* or *investigators*. The leads are in charge of the working group or the project, and the investigator, mainly the principal investigator (PI), is in charge of collecting the data from a cohort. Within each working group, there are *projects* that work on a more specific aspect. For example, within the Schizophrenia Working Group, there is a project on *Cannabis effects on structural measures in SZ*. Each project collects data from multiple cohorts and analyzes the data to reach a conclusion. Each project has many members and some may have leadership roles. Working groups last a long time, while projects last a finite time and results in a publication. A *cohort* is a dataset collected by a PI, consisting of imaging and/or genomics data for a set of individuals. A cohort can be part of multiple working groups, and a project can use cohorts from various working groups.

Sample	Criteria for Inclusion/Exclusion
Amsterdam	<p>Inclusion criteria for BD-I/II subjects were: age 18-60 years; age at first mood episode ≤ 40 years; a history of at least two Major Depressive Episodes (MDEs); current MDE or remission (i.e. not fulfilling the criteria of MDE or (hypo)manic episode); and illness duration of ≥ 5 years since the first episode. Exclusion criteria for BD subjects were: current (hypo)mania; a current comorbid axis-I disorder (including any current substance use disorder; assessed by SCID-I) except for anxiety disorders; a diagnosis of cluster B personality disorder (assessed by SCID-II, if suspected based on (hetero)anamnestic information); current use of antidepressants, anticonvulsants, mood-stabilizers or antipsychotics (stopped at least 1 month before scanning); electroconvulsive therapy within 2 months before scanning. Incidental benzodiazepine use was allowed, but was stopped at least 1 day before scanning. BD subjects were excluded if (hypo)manic episodes solely appeared during the use of antidepressants, again to ensure inclusion of 'true' BD subjects; and in case of a Young Mania Rating Scale (YMRS) score > 8, to prevent inclusion of BD subjects with (subclinical) (hypo)manic symptoms. Inclusion criterion for HC was age 18-60 years. Exclusion criteria for HC were: presence of a lifetime psychiatric diagnosis (Axis I); first-degree relatives with a history of a psychiatric diagnosis (Axis I); use of any psychopharmacological agent.</p> <p>Furthermore, for all subjects contra-indications for MRI-scanning led to exclusion: a history of head trauma or neurological disease; severe general physical illness; claustrophobia or implanted metal objects.</p>

Figure 3: Inclusion and exclusion criteria for the Amsterdam cohort used in [Hibar et al 2017].

In ENIGMA, when a project conducts a study, they may not use the entire cohort that a PI has. Instead, they may use a subset of the cohort and only use data from some of the participants. This subset of the cohort is called a *project cohort*. In a working group, there are many members that work for that working group. These members can also work in other working groups. A member may have a certain role within a working group. For example, they can be a lead or a contributor. Multiple people can have the same role. The same situation applies to projects. The data that is collected in a given cohort is retrieved by following a certain protocol. Since the *image acquisition protocol* is different for each cohort, each cohort has a unique protocol.

The *inclusion and exclusion criteria* for the cohorts and the details for the acquisition protocol also had to be represented. The inclusion and exclusion criteria determine whether or not someone can be a participant of a cohort. Figure 3 gives an example of the inclusion/exclusion criteria found in [Hibar et al. 2017]. The inclusion criteria are the criteria that someone must meet in order for them to be a participant. An example of an inclusion criteria in that article is that the age must be in the range of 18 to 60 years. In order for someone to be included in the cohort, they must be within this age range. The exclusion criteria are what renders someone ineligible to be part of the cohort. An example of an exclusion criteria in that article is the current use of antidepressants. Thus, if someone met all of the inclusion criteria but used antidepressants, they would not be allowed to be a participant.

Another important piece of information about a cohort is the experimental protocol used to acquire imaging or genomic data. The image acquisition protocol provides the details of how a brain scan was collected. There are different types of brain scans and different ways of acquiring that data with a scanner. Several different sequences can be used. Figure 4 provides a small section of the image acquisition table from [Hibar et al. 2017]. It is important to note that the following table is rather simple compared to the many different image acquisition tables.

Site	Sequence	Field Strength	Acquisition Direction	# of Slices	Slice Gap	Voxel Size (mm3)	TI	TE	TR	Flip Angle	Citation	Segmentation
Amsterdam	3D T1-weighted turbo field echo (TFE)	3T scanner Philips Gyroscan Intera	coronal	182	0mm	1x1x1.2	0ms	4.6ms	9.621ms	8	1, 2	FreeSurfer (5.0)

Figure 4: Image acquisition protocol table for the Amsterdam cohort used in [Hibar et al 2017].

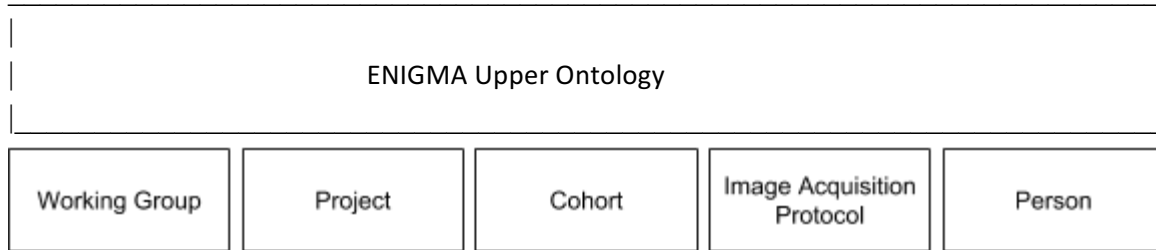


Figure 5: Main modules of the ENIGMA ontology.

The properties added to the ontologies needed to represent information that appears in published papers of ENIGMA. In published papers, there are usually a demographics table that lists the important details of the cohorts, such as who the principal investigators are or how many participants were part of the cohort (as shown in Figure 3). Other tables show the common properties of the image acquisition protocols, such as the acquisition direction or voxel size (as shown in Figure 4).

3. DESIGN OF THE ONTOLOGY

3.1 Design for Modularity

For the ENIGMA ontology, a modular design was used. Figure 5 shows the main modules. The main reason behind this decision was that there were many details that needed to describe each concept within ENIGMA. By having different modules, each representing different entities, it is easier to manage all the information. The ENIGMA ontology has six different modules. There is an upper ontology that describes the basic relationships between the various core concepts within ENIGMA and gives a general overview of the entire ontology. The other five modules represent the working groups, the projects, the cohorts, the image acquisition protocols, and the people. Each module only contains properties of that concept or properties that refer to that concept. For example, within the cohort module, there are properties such as “hasTreatment.” Since “hasTreatment” is one of the inclusion criteria of the cohort, it is a property that will be seen in the cohort ontology and will not appear in any of the other modules.

3.2 Design for Usability

Some properties were added to allow for easier user interaction with the ENGIMA-ODS framework. For example, a property was necessary to represent the age range of a cohort. Initially, the property was created as an object property that had a range of `AgeRange`. Each age range would then have two properties that represented the minimum and maximum age for the range. To input the data, a user would have to create a data structure for the age range and edit the minimum and maximum values. Thus, the age range was represented simply as two data properties of a cohort: `hasMinAge` and `hasMaxAge` to eliminate the need of a separate data structure.

In order to make the system easier to use, some properties were inferred. There were two ways that this task was accomplished. The first way was to make certain properties as inverses of one another. For example, `hasPI` and `isPIOf` are inverses of one another. By creating inverse properties, certain information on the system will be automatically filled out when the inverse property is used. Thus, the users will not have to manually add all of this information. Another way that certain properties were inferred was through subproperties. If a subproperty is created, the parent property is automatically inferred. An example would be the `hasWGMember` property and the `hasSeniorLeadOfWG` property. If John Doe is the senior lead of a working group, he will automatically be a member of that working group as well. Users will not have to enter their name twice.

Having subproperties also supports showing information selectively. For example, in a person page, researchers wanted to see which projects someone contributed to but did not want to see their specific role. So, they wanted to see that John Doe contributed towards a specific bipolar project but did not want to see whether or not he was a senior lead. Thus, by making certain properties into subproperties, the extent of the details shown can be adjusted.

Lastly, enumerated data properties were added to make it quicker for users to input information. This means that the property specifies a finite list of possible values. For example, in the case of study design, the only possible choices in ENIGMA are `population-based` or `case-control`. By limiting the values, organizing the information within the wiki will be a lot easier than if the value was specified as a string. If it was simply a string, then the system would accept `Case control` and `case-control` as two different study designs.

3.3 Design of Collections

One of the hardest concepts to represent semantically was the cohort as a collection of individuals. As mentioned above, a project may not use the entire cohort that a PI has, and instead, use a subset of the cohort called the project cohort. Project cohorts were made a subclass of the cohort class. Although a project cohort and a cohort both have the same properties, the researchers have noted that users of the wiki will not need to know certain details for the cohort

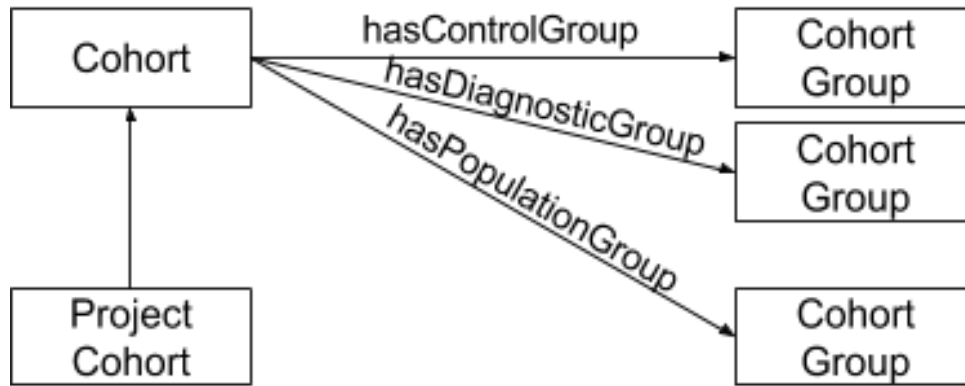


Figure 6: An illustration of the relationships between a cohort, project cohort, and the cohort groups.

but will need to know the details for the project cohort. For example, the median or mean age of the participants are noted on the publication that results from an ENIGMA project. Although the cohort may have this information, the users will not need to look up this information. However, they will need it. Thus, some properties are needed for the project cohorts while they are defined as general cohort properties. Since the project cohort is a subclass of the cohort, it will inherit these cohort properties.

Another modeling issue for the cohorts were differentiating between the healthy controls and the patients. In ENIGMA, a cohort can be either a population-based cohort or a case-control cohort. A population-based cohort includes a general group of participants. An example of a population-based cohort would be a cohort that includes all of the students attending a school. A case-control cohort, on the other hand, only scans people based on a certain characteristic that they have. For example, if patients were scanned because they had bipolar disorder, these patients will be the cases of the case-control cohort and be part of the diagnostic group. In a case-control cohort, there may also be a group of participants that were chosen to be part of the control group because they did not have those characteristics. In ENIGMA papers, the inclusion and exclusion criteria for the control participants and the diagnostic participants differ. For example, in Figure 3, the controls were excluded if they had a first degree relative with a history of a psychiatric disorder, whereas this was not an exclusion criteria for the patients of the diagnostic group. To deal with this discrepancy, another class called CohortGroup was added. The CohortGroup is not a subclass of Cohort because it does not have the same properties as a Cohort. A CohortGroup is either the diagnostic group, control group, or population group of a cohort. A CohortGroup has properties that describe the inclusion and exclusion criteria for that specific group. A property relates the CohortGroup to the Cohort so that the users will know which group of people is the control, patient, and population group. The relationships between the cohort, the project cohort, and the cohort groups can be seen in Figure 6.

Extra information	
Property	Value

Figure 7: The extra information section allows users to create new properties.

Lastly, the cohort modeling also posed a problem regarding completion status. A cohort and project can either be finished or be ongoing. Depending on the completion status of the cohort, the properties that the principal investigator has information for are different. The difficulty with this issue is that even though a complete project cohort has slightly different properties from an ongoing project cohort, they are both project cohorts. To deal with this issue, all of the properties that apply to project cohorts that are complete and all those that apply to the project cohorts that were ongoing were added to the ontology. The principal investigator or lead who is uploading the information will use their judgement to decide which properties to fill out and which ones to leave empty. There is an additional property that states the status of the project or cohort. Therefore, users will know that the information that is presented is describing a project cohort with a certain completion status.

3.4 Design for Simplicity

There is no standardized vocabulary for the inclusion/exclusion criteria and the image acquisition protocols. Their descriptions can get extremely detailed. For example, an inclusion criteria may be to have second psychotic episode before the age of 40. Such detail would make an ontology very complex and although detailed, the information will not be organized in an efficient manner. Thus, the paragraphs describing the inclusion and exclusion criteria and the image acquisition protocol had to be simplified, and the main properties had to be identified. In the descriptions for the inclusion and exclusion criteria, the most common phrases were turned into properties, such as “hasPersonalHistoryOf,” “hasTreatment,” or “doesNotIncludeIntellectualDisability.” In the Cohort ontology, users can see that there are properties that are specific to the inclusion criteria and properties that are specific to the exclusion criteria. The naming convention used specifies that all properties that have “doesNot” or “excludes” in the beginning are exclusion criteria. Since the inclusion and exclusion criteria are very detailed, there will always be properties that users want to define that are not part of the ontology. Users can create these additional properties through the ENGIMA-ODS interface as seen in Figure 7. This capability is explained in detail in [Gil et al 2017].

A complex aspect of inclusion and exclusion criteria was describing the details of the disorder, treatment, condition, and medical rating scale. For example, the age of onset of a disorder may be important because the researchers might want to study an early onset disease. For treatment and condition, they need to know how recently they were done. For medical rating scale, they may need the participants to be within a certain interval. To fix this problem, four

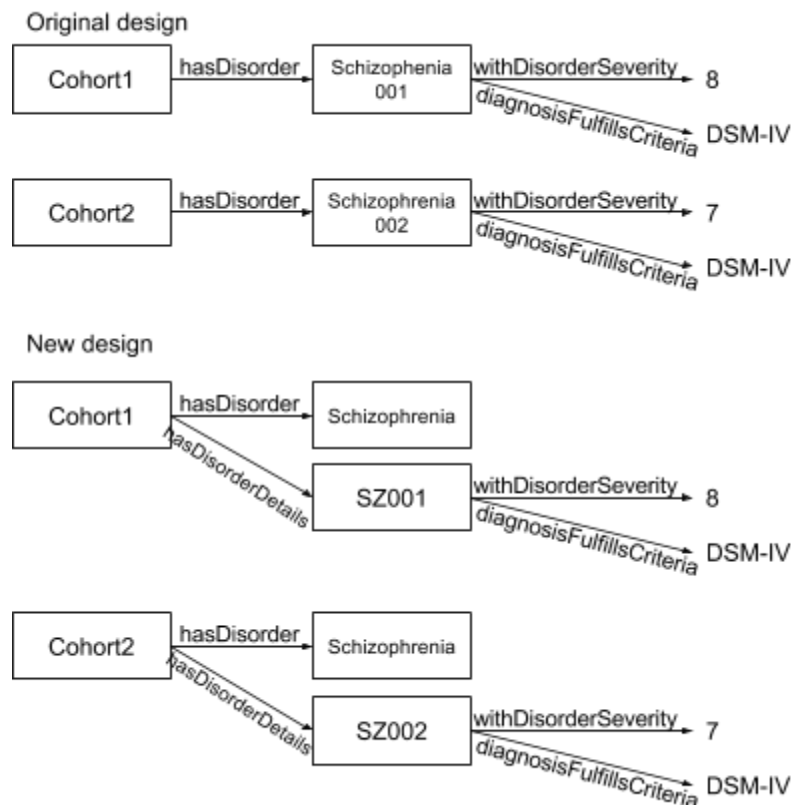


Figure 8: Depiction of the modeling differences between the two designs.

Acquisition Information
Data acquisition for the AddNeuroMed study was designed to be compatible with that of ADNI. The imaging protocol for both studies included a high-resolution sagittal 3D T1-weighted MPRAGE volume (voxel size 1.1 × 1.1 × 1.2 mm ³) and axial proton density/T2-weighted fast spin echo images. The MPRAGE volume was acquired using a customized pulse sequence specifically designed for the ADNI study to ensure compatibility across scanners. Full brain and skull coverage was required and detailed quality control was carried out on all MRI data according to the AddNeuroMed quality control procedure [Simmons et al 2009, Simmons et al 2011]

Figure 9: Image acquisition protocol described through a paragraph in [Hibar et al. 2015].

classes were created, each class describes the specific details of disorder, treatment, condition, and medical rating scale. The cohort group will have a disorder as well as an instance of a disorder details. There are two design options: one is to consider this information to be part of the definition of the disease, or for them to be details of this disorder. The difference between the two designs is depicted below in Figure 8. The second design was chosen to enable disease queries.

SIEMENS MAGNETOM Avanto syngo MR B15

\\USER\INVESTIGATORS\Ravi\ADNI-phantom\QC Phantom-Localizer			
TA: 0:10	PAT: Off	Voxel size: 1.9×1.5×8.0 mm	Rel. SNR: 1.00 SIEMENS: gre

Properties			
Prio Recon	Off	Base resolution	192
Before measurement		Phase resolution	75 %
After measurement		Phase partial Fourier	Off
Load to viewer	On	Interpolation	Off
Inline movie	Off	PAT mode	None
Auto store images	On	Matrix Coil Mode	Auto (CP)
Load to stamp segments	Off	Distortion Corr.	Off
Load images to graphic segments	Off	Prescan Normalize	Off
Auto open inline display	Off	Normalize	On
AutoAlign Spine	Off	Raw filter	On
Start measurement without further preparation	Off	Intensity	Weak
Wait for user to start	Off	Slope	25
Start measurements	single	Elliptical filter	Off

Routine		Geometry	
Slice group 1		Multi-slice mode	Sequential
Slices	1	Series	Ascending
Dist. factor	20 %	Saturation mode	Standard
Position	Isocenter	Special sat.	None
Orientation	Sagittal	Set-n-Go Protocol	Off
Phase enc. dir.	A >> P	Table position	H
Rotation	0.00 deg	Table position	0 mm
Slice group 2		Inline Composing	Off
Slices	1	System	
Dist. factor	20 %	Body	Off
Position	Isocenter	HE1	On
Orientation	Coronal	HE3	On
Phase enc. dir.	R >> L	HE2	On
Rotation	0.00 deg	HE4	On
Slice group 3		Positioning mode	REF
Slices	1	MSMA	S - C - T
Dist. factor	20 %	Sagittal	R >> L
Position	Isocenter	Coronal	A >> P
Orientation	Transversal	Transversal	E << L

Figure 10: An example of an image acquisition protocol from ADNI described through a template [ADNI 2017].

There was a similar issue with the image acquisition protocols. The image acquisition protocol is described in the paper either in a paragraph, table, or template. An example of an image acquisition protocol described in paragraph form can be seen below in Figure 9 and is from [Hibar et al. 2015]. An example of the image acquisition protocol displayed in table form can be seen in Figure 4. A standard image acquisition protocol for Alzheimer’s Disease Neuroimaging Initiative that is written on a template can be seen in Figure 10.

Since paragraphs are not very uniform, key details from the paragraphs describing the image acquisition protocols had to be selected. In addition, templates are so diverse that they do not provide a clear idea into which properties are necessary. There are multiple template formats

that are used to describe acquisition protocols. Since there are various different templates, the descriptions are different depending on which format was used. Simplifying this information was necessary to reduce the image acquisition protocols to the ones that were the most common. The process used to find the most common properties were similar to the process used for simplifying the inclusion/exclusion criteria. The common phrases and characteristics that appeared on various tables and papers were included in the ontology. The extremely detailed properties were excluded as they only pertained to a very few number of acquisition protocols and was not efficient to include.

4. DISCUSSION

The ontology was used to populate the ENGIMA-ODS framework. Figure 11 shows a snapshot of a summary table generated by the system. It shows the projects of the SZ Working Group on the left column, and the rows show their descriptions, the junior and senior leads, and their contact person. Each of the group and person names is hyperlinked to a page that describes their properties, and someone with the valid permissions can edit the information and add values for the properties. Creating a new instance creates a new page, and one can easily select a class for the page by selecting a category (A category is the wiki terminology for an ontology class). It is easy to navigate across pages. One can click on an image acquisition protocol from the cohort page. This will lead to another page that has all the properties for the image acquisition protocol. Through the interface, all the relationships among entities can easily be identified. [Jang et al 2017] describe in more detail the ENIGMA-ODS framework.

There are many properties that are currently on the ENIGMA-ODS framework, especially for cohorts. Currently, when one tries to edit these pages, the properties are not listed in an organized fashion. Thus, it may take some time finding a specific property. It would be a lot cleaner and more user-friendly if all the properties for a certain category were organized into groups

The information system for ENIGMA is still a work in progress. Even at its early stages, the system shows a lot of potential. In addition to being a system that researchers and collaborators can use to keep track of their information and data and query information easily, the information system can also be used to automatically generate selected portions of scientific papers. We are working on generating the author list and various tables for project papers using the information in the system. These automatically generated portions will make writing papers a far simpler task as they can be created by calling certain properties within the ENIGMA-ODS framework.

The table below summarizes all the projects sponsored by the Schizophrenia Working Group:

	HasPBriefDescription (L)	HasPJuniiorLead (L)	HasPSeniorLead (L)	HasContactPerson (L)
SZCorticol	This was the second prospective meta-analysis of the full group and is currently being written up, but new groups are encouraged to continue adding to the results, and additional analyses will be developed.	Sinead Kelly Sara M Gary Y	Nailin Yao Gary Y Fude Yang Esther Walton Jian Wenhao Theo G. M. vanErp Jessica Turner	Jessica Turner
SZNegativeSymptoms	examines relationship between left orbitofrontal thickness and negative symptoms	Esther Walton	Stephan Erlich	Esther Walton Stephan Erlich
SZPositiveSymptoms	examines relationship between superior temporal gyrus thickness and positive symptoms	Esther Walton	Stephan Erlich	Stephan Erlich Esther Walton

Figure 11: Snapshot of the ENIGMA-ODS interface.

Since the ENIGMA-ODS framework is only in its beginning stages, there are some functionalities that may be necessary to add in the future. There will be a need to keep certain information within the wiki from the public. Project leads and PIs should be able to give or restrict access to people to certain pages and control who has the privileges to edit or view certain pages.

Another extension that can be added is by creating the ability to mass populate the system. Currently, all of the properties have to be manually entered, a process which is not very efficient. The ideal solution would be to support uploading of spreadsheets onto the system then populate the system with the information.

Mapping to existing vocabularies would also be helpful. There are certain concepts for which ontologies already exist. For example, the disorder of inclusion and exclusion criteria of a cohort. Although ENIGMA focuses mainly on neurological and mental disorders, set criteria often mention other types of diseases. Incorporating existing ontologies such as the Human Disease Ontology (DOID) [Osborne et al 2009; DOID 2017] to describe diseases will increase the usability of the system.

5. RELATED WORK

There are various methods for collaborative knowledge collection, one of which is through a semantic wiki [Gil 2013]. A few semantic wiki platforms do support editing collaboratively [Bry et al. 2012]. One such platform is the Organic Data Science (ODS) platform

[Gil et al. 2017]. Currently, ENIGMA is not the only project using the Organic Data Science platform. LinkedEarth [Gil et al. 2017] was also built on the Organic Data Science platform and focuses on representing data for the paleoclimate community. Since LinkedEarth uses the same platform as ENIGMA, many of the functionalities present for LinkedEarth are also.

Another similar knowledge network is DCvO [Ma et al 2017]. Like the ENIGMA information system, DCvO provides a way to describe important entities of the communities they represent. DCvO represents relationships projects, people, affiliations, and data types. Since the DCvO is a network for deep carbon science, it has categories and properties that do not relate to ENIGMA as ENIGMA is a neuroscience network. Like ENIGMA, DCvO strives to enable researchers to record information regarding data processing, funding, publications, and more. The faceted browsers of DCvO allow for specific queries, as does the ENIGMA system.

7. CONCLUSIONS

This report describes the development of an ontology for the ENIGMA neuroscience collaboration. There is a lot of complex information within ENIGMA groups but no uniform way of representing that information. To solve this problem, a core ontology was built on the Organic Data Science platform. The ontology contains many concepts and properties to represent ENIGMA entity. The main concepts are organized into separate modules, and the properties describe concepts and also establish links among them. The ontology represents working groups, project, people, and cohorts, as well as inclusion/exclusion criteria and imaging acquisition protocols. The challenges of representing complex information, usability, and collections led to various modeling decisions. The ontology presented provides a way for researchers to make queries easily, to find potential collaborators, projects, and datasets.

Acknowledgments. This work was supported by a grant from the Kavli Foundation and by the National Science Foundation with award ICER-1541029. The ontology was developed using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health. The author would like to thank Neda Jahanshad and Agnes McMahon for explaining the ENIGMA collaboration, Daniel Garijo and Yolanda Gil for their guidance in understanding how to develop ontologies, and other members of the ENIGMA-ODS project for their feedback on this work.

REFERENCES

[ADNI 2017] Alzheimer’s Disease Neuroimaging Initiative (ADNI). MRI Scanner Protocols. Available from <http://adni.loni.usc.edu/methods/documents/mri-protocols/>. Last accessed 09 August 2017.

- [DOID 2017] Human Disease Ontology. Available from BioPortal at <http://bioportal.bioontology.org/ontologies/DOID>. Last accessed 09 August 2017.
- [Gil et al. 2017] A Controlled Crowdsourcing Approach to Scientific Ontology Development and Data Annotation. Gil Y, Garijo D, Ratnakar V, Khider D, Emile-Geay J, and McKay N. Proceedings of the Sixteenth International Semantic Web Conference (ISWC), Vienna, Austria, 2017
- [Gil 2013] Social Knowledge Collection. Gil, Y. In Handbook of Human Computation, 2013. doi:10.1007/978-1-4614-8806-4_24
- [Hibar et al. 2015] Common genetic variants influence human subcortical brain structures. Hibar, DP, Stein JL, Renteria ME, et al. Nature 520, 224–229, 2015.
- [Hibar et al. 2017] Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. Hibar DP, Westlye LT, Doan NT, Jahanshad N, et al. Molecular Psychiatry (2017). doi:10.1038/mp.2017.73
- [Jang et al 2017] Towards Automatic Generation of Portions of Scientific Papers for Large Multi-Institutional Collaborations Based on Semantic Metadata. Jang, M, Patted, T, Gil, Y, Garijo, D, Ratnakar, V, Ji, J, Wang, P, McMahon, A, Thompson, P. M, and N. Jahanshad. Proceedings of the Workshop on Enabling Open Semantic Science, co-located with the Sixteenth International Semantic Web Conference (ISWC), 2017.
- [Ma et al. 2017] Weaving a Knowledge Network for Deep Carbon Science. Ma X, West P, Zednik S, Erickson J, Eleish A, Chen Y, Wang H, Zhong H, Fox P. Frontiers in Earth Science 5, 2017.
- [Musen 2015] The Protégé project: A look back and a look forward. Musen, MA. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.
- [Osborne et al 2009] Annotating the human genome with Disease Ontology. Osborne JD, Flatow J, Holko M, Lin SM, Kibbe WA, Zhu LJ, Danila MI, Feng G, Chisholm RL. BMC Genomics 10 S1:S6, 2009.
- [Protégé 2017] The Protégé Ontology Editor. Version 5.2.0. Accessible from <https://protege.stanford.edu/>. Last accessed 09 August 2017.
- [Thompson et al 2014] The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, Toro R, Jahanshad N, et al. Brain Imaging and Behavior, 8(2), 2014.