An Extensible Ontology Modeling Approach Using Post Coordinated Expressions for Semantic Provenance in Biomedical Research

Joshua Valdez[1], Michael Rueschman[2], Matthew Kim[2], Sara Arabyarmohammadi[1], Susan Redline[2], and Satya S. Sahoo[1(✉)]

[1] Institute for Computational Biology and Electrical Engineering and Computer Science Department, Case Western Reserve University, Cleveland, OH, USA
{joshua.valdez, sara.arabyarmohammadi, satya.sahoo}@case.edu

[2] Department of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard University, Boston, MA, USA
{mrueschman, mikim, sredlinel}@bwh.harvard.edu

Abstract. Provenance metadata describing the source or origin of data is critical to verify and validate results of scientific experiments. Indeed, reproducibility of scientific studies is rapidly gaining significant attention in the research community, for example biomedical and healthcare research. To address this challenge in the biomedical research domain, we have developed the Provenance for Clinical and Healthcare Research (ProvCaRe) using World Wide Web Consortium (W3C) PROV specifications, including the PROV Ontology (PROV-O). In the ProvCaRe project, we are extending PROV-O to create a formal model of provenance information that is necessary for scientific reproducibility and replication in biomedical research. However, there are several challenges associated with the development of the ProvCaRe ontology, including: (1) *Ontology engineering*: modeling all biomedical provenance-related terms in an ontology has undefined scope and is not feasible before the release of the ontology; (2) *Redundancy*: there are a large number of existing biomedical ontologies that already model relevant biomedical terms; and (3) *Ontology maintenance*: adding or deleting terms from a large ontology is error prone and it will be difficult to maintain the ontology over time. Therefore, in contrast to modeling all classes and properties in an ontology before deployment (also called *precoordination*), we propose the "ProvCaRe Compositional Grammar Syntax" to model ontology classes *on-demand* (also called *postcoordination*). The compositional grammar syntax allows us to re-use existing biomedical ontology classes and compose provenance-specific terms that extend PROV-O classes and properties. We demonstrate the application of this approach in the ProvCaRe ontology and the use of the ontology in the development of the ProvCaRe knowledgebase that consists of more than 38 million provenance triples automatically extracted from 384,802 published research articles using a text processing workflow.

H. Panetto et al. (Eds.): OTM 2017 Conferences, Part II, LNCS 10574, pp. 337–352, 2017.
https://doi.org/10.1007/978-3-319-69459-7_23

Keywords: Precoordinated and postcoordinated expression · Ontology engineering · Provenance metadata · W3C PROV specification · ProvCaRe semantic provenance

1 Introduction

Scientific reproducibility is critical for ensuring validation of research results, scientific fidelity, and enabling the advancement of science through rigorous design of experiments [1, 2]. Therefore, the increasing adoption of data-driven research techniques, for example use of Big data in biomedical and healthcare research for better understanding of disease mechanism and drug discovery, has led to greater focus on scientific reproducibility [3, 4]. Multiple guidelines and best practices have been developed to ensure transparent reporting of scientific results that can be successfully replicated. For example, the US National Institutes of Health (NIH) has announced the "Reproducibility and Rigor" guidelines that requires biomedical researchers to provide contextual information for transparent reporting of research studies [5]. This contextual information describing the origin or source of data is called provenance metadata. Provenance metadata has been extensively studied in computer science for reproducibility in workflow systems and tracing data in relational database systems [6–8]. By leveraging provenance metadata in biomedical and healthcare research, researchers will have improved ability to collaborate, share data, identify "best practices" for reproducible scientific research [9]. In addition to scientific reproducibility, provenance metadata is also essential for evaluating data quality and computing trust value [10, 11]. Due to its application in a variety of domains, provenance has been modeled using multiple approaches, for example the Open Provenance Model (OPM) represented causal relationship between different provenance terms [12]. Similarly, the Provenir ontology used Semantic Web technologies, including the Web Ontology Language (OWL) [13] to incorporate partonomy, causal, transformation, and other categories of relationships to accurately represent provenance metadata [14].

The World Wide Web Consortium (W3C) provenance working group used various properties of these provenance modeling approaches to define the PROV specifications as a common representation model in 2013 [15]. The W3C PROV specifications consist of the PROV Data Model (PROV-DM), [15] a formal representation of the data model using description logic-based Web Ontology Language (OWL2) called PROV Ontology (PROV-O) [16], and a set of constraints to define "valid" provenance representations called PROV Constraints [17]. In particular, the W3C PROV Ontology was developed as an upper-level reference ontology that can be extended to model domain-specific provenance terms while ensuring interoperability through the use of a common set of ontology classes and properties [16]. We have extended the W3C PROV specifications to define a new provenance framework called Provenance for Clinical and Healthcare (ProvCaRe) to support scientific reproducibility in biomedical and healthcare domain. The ProvCaRe framework defines a reference model consisting of three categories of provenance metadata terms that we have identified as necessary for scientific reproducibility in biomedical research:

1. **Study Method:** The design of the research study in terms of study design, sampling, randomization technique and interventions (in experiments), data collection approach, and data analysis techniques (e.g., statistical models) are examples of provenance metadata describing Study Method;
2. **Study Tools:** The different instruments and their parameter values that are used to record and analyze data in a research study is the second essential component of the ProvCaRe framework. For example, the strength of the magnet used in a Magnetic Resonance Imaging (MRI) instrument is important provenance information that will allow other researchers to use an equivalent MRI machine to replicate the findings of the original experiment; and
3. **Study Data:** The provenance metadata describing the contextual information about the data elements used in a scientific experiment, for example drug information, demography information of participants, and timestamp values, are necessary to allow other researchers to replicate a given experiment.

Given the vast scope of biomedical and healthcare research, we initiated the development of the ProvCaRe framework using sleep medicine research as an example domain to identify, extract and analyze provenance information associated with research studies. We are using data from one of the largest repositories of sleep medicine studies at the National Sleep Research Resource (NSRR), which is working to release data from more than 40,000 sleep studies collected from 36,000 participants [18]. The NSRR project is an example of biomedical Big Data and it aims to allow researchers to validate results of previous studies using larger datasets from greater number of research studies and also facilitate the development of data-driven techniques in sleep medicine. Therefore, the systematic characterization of provenance metadata describing the research studies that involve analysis of NSRR datasets will not only demonstrate the role of provenance in scientific reproducibility, but also demonstrate the scalability of the ProvCaRe framework. Our objective is to model the provenance metadata extracted from NSRR related research studies using the "`subject → predicate → object`" triple model of W3C Resource Description Framework (RDF) [19]. The provenance terms used to construct the RDF provenance graphs are being modeled in the ProvCaRe ontology, which extends the W3C PROV Ontology and the resulting provenance graphs also conform to the PROV specifications [16].

The W3C PROV Ontology consists of three "core" classes, namely `prov:Entity`[1], `prov:Activity`, and `prov:Agent`, with nine "core" object properties, namely `prov:wasGeneratedBy`, `prov:wasDerivedFrom`, `prov:wasAttributedTo`, `prov:startedAtTime`, `prov:used`, `prov:wasInformedBy`, `prov:endedAtTime`, `prov:wasAssociatedWith`, and `prov:actedOnBehalfOf` [16]. Figure 1 shows the PROV-O schema with an illustrative representation of provenance metadata in sleep medicine research. The `Entity` class represents any physical, digital, or conceptual information resource. The `Activity` class represents information resources that occur over a period of time and Agent class represents information resources that are associated with `Activity`, `Entity` or have some responsibility related to another `Agent`. The object properties are used to link the

[1] `prov` represents the http://www.w3.org/ns/prov#namespace.

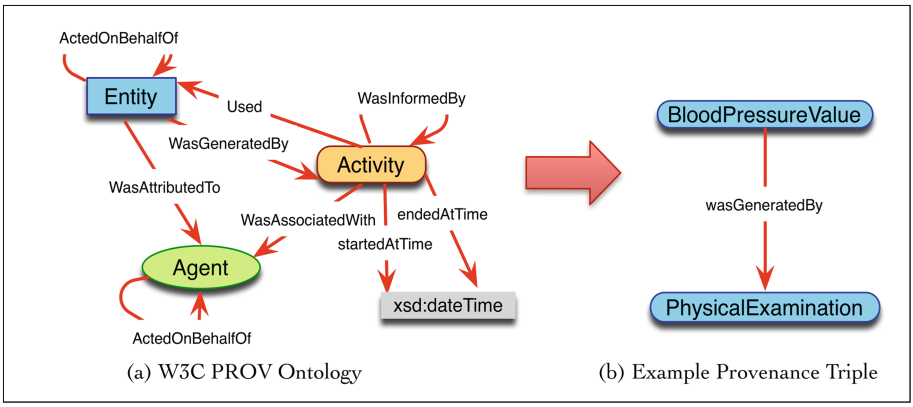

Fig. 1. (a) The three primary classes and object properties of the W3C PROV Ontology (PROV-O). (b) An example showing representation of provenance information using RDF subject, predicate, and object structure.

provenance terms, for example blood pressure value (`Entity`) was generated during (`wasGeneratedBy`) a physical exam of the patient (`Activity`). These "core" classes and properties together with other PROV Ontology terms (as described in the PROV-O specifications [16]) are being extended to model biomedical domain-specific provenance information in the ProvCaRe framework. The use of PROV-O as the upper-level ontology will facilitate interoperability among provenance applications that conform to the PROV specifications. However, a key challenge for the ProvCaRe ontology is ensuring comprehensive coverage of the potentially hundreds of thousands of biomedical domain-specific terms in a single provenance ontology using *precoordinated* class expressions. Precoordinated class expressions are ontology constructs that already "built-in" in an ontology before the ontology is deployed or used (a detailed description of precoordination is presented in work by Rector et al. in [20]).

The biomedical domain covers a wide range of disciplines, including respiratory disease, neurology, and cardiovascular research, among others, and therefore it is almost impossible for a single ontology to model the relevant terms with consistency in a reasonable amount of time. In addition, there are more than 500 biomedical ontologies already available from the National Center for Biomedical Ontologies (NCBO) that represent a variety of biomedical terms at different levels of granularity and detail [21]. For example, the Human Phenotype Ontology (HPO) models abnormal phenotypes in human diseases and it covers different aspects of these abnormalities, including the mechanism for inheritance of these abnormalities, their onset and clinical course, and different categories of the abnormalities [22]. HPO includes more than 10,000 classes with many of the classes mapped to other biomedical ontologies. Similarly, the Systematize Nomenclature of Medicine Clinical Terms (SNOMED CT) is being developed as a comprehensive ontology for diseases and phenotypes with a large number of clinical terms modeled in the ontology [23]. The US edition of SNOMED CT 2015 version includes 300,000 concepts with more than 103,000 classes representing clinical findings. Therefore, it is intuitive to re-use these large numbers of

existing ontology classes in the ProvCaRe project to model domain-specific terminology instead of re-creating the terms in the ProvCaRe ontology. The re-use of existing ontology classes also conforms to the ontology engineering best practice and facilitates interoperability across ontology-driven applications [24].

2 Background and Related Work: Use of Postcoordination in Biomedical Ontologies

Formal modeling of attributes related to the design and template of clinical and basic research studies have led to the development of multiple ontologies in biomedical research. The Ontology for Clinical Research (OCRe) was developed as part of a comprehensive effort to model protocols used in clinical research studies, including a classification of study designs, the plan components of the study protocols, and concepts describing statistical data analysis methods [25]. The OCRe project defines multiple attributes to represent various aspects of a research study, including the sampling method, number of participant groups, and whether a study is a longitudinal cohort or cross-sectional study. The OCRe project also developed a data annotation approach called Eligibility Rule Grammar and Ontology (ERGO), which extracts structured information regarding eligibility criteria used to identify participants in research studies [26]. A formal description of eligibility criteria is important metadata information that can be used by other researchers to replicate a biomedical or healthcare study. Similar to OCRe, the Ontology for Biomedical Investigations (OBI) models various attributes of basic sciences experiments, for example in genomics, proteomics, and parasite research domains [27]. The classes in OBI broadly represent five categories of entities, including the objects used in experiment called material entity, activities in experiments such as planned processes, the data related to an experiment called information entities, different roles of participants in experiments, and instruments. OBI has been used in annotation of multiple biomedical databases, for example the Eukaryotic Pathogen Database and the Immune Epitope Database.

In contrast to OCRe and OBI, SNOMED CT is a model of clinical terminology organized into 19 top level concepts, for example clinical findings, procedure, specimen, and body structure. These terms are linked to each other using attributes or properties, for example causative agent, associated morphology, and finding site. To address the challenging requirement of modeling extremely large variety of concepts from different biomedical disciplines, SNOMED CT uses two approaches to represent terms: (1) *Precoordinated Expressions*, and (2) *Postcoordinated Expressions*. Precoordinated expressions in SNOMED CT consist of a single class and are modeled in one of the 19 class hierarchies, for example `Sleep disorder` (ID: C0851578) is modeled as a subclass in the hierarchy of the top-level concept `Clinical finding` (ID: C0037088). However, it is almost impossible to model all possible attributes of a disease, which may evolve as new biomedical discoveries are made or additional clinical details that were not considered before and need to be modeled in context of a specific application. To address this challenge, SNOMED CT uses post coordination expressions to represent new terms by combining more than one SNOMED CT term using a set of rules defined in the SNOMED CT compositional grammar specification

[23]. For example, the post coordinated expression | hip joint | : | laterality | = | right | represents right hip joint using the classes hip joint and right together with the property laterality. We propose to use a similar approach to model provenance information for the biomedical domain through creation of postcoordinated expressions using classes from the ProvCaRe ontology together with classes from existing biomedical ontologies. We describe the details of our approach in the next section.

3 Modeling Provenance Metadata in ProvCaRe Ontology Using Postcoordinated Expressions

The ProvCaRe framework models the provenance description of a scientific study that may enhance the ability to replicate the study by researchers in other institutions or groups. The three objectives of the ProvCaRe framework are: (1) Create a biomedical domain-specific provenance ontology, (2) Extract provenance metadata from published biomedical articles and generate provenance graphs for analysis, and (3) Develop a provenance knowledgebase for users to search and identify research studies that can be replicated to validate important results or design new experiment studies. The three components of the ProvCaRe framework, namely Study method, data, and tools, were developed in close collaboration with a data manager working on the NSRR project. A data manager is responsible for working with researchers to identify the data needed to replicate results from previous studies and extract data for new research studies. Therefore, they are ideally placed to identify provenance information required for scientific reproducibility.

As described in Sect. 1, the three components of the ProvCaRe framework represents three essential provenance information types corresponding to the method used to conduct the research study (*Study Method*), the data used in the study as well as results generated from the study (*Study Data*), and details of the instruments used in the study (*Study Instrument*). The provenance information corresponding to these three components can be modeled in detail, which is important to accurately capture the contextual information necessary for replicating previous studies. For example, the Study Method term can be further subdivided into three categories of: (a) Study Design, (b) Study Data Collection Method, and (c) Data Analysis Method. Similarly, the Data Analysis Method can be further extended to model various categories of statistical data analysis methods, for example inferential or descriptive statistics. We use sleep medicine as an example domain with data from the NSRR project to define the ProvCaRe postcoordination-based ontology modeling approach and demonstrate the applicability of the ProvCaRe ontology.

3.1 Provenance Metadata in Sleep Medicine Research

NSRR is the largest repository of publicly available research sleep medicine data and it includes data from research studies that are representative of a wide variety of topics, for example cardiovascular diseases, neurocognitive functions, and metabolic disorders related to sleep disorders. Therefore, it is well suited to develop the ProvCaRe

framework in terms of scalability and it is representative of the complexity of the biomedical research domain. Biomedical research studies are often defined in terms of the well-known *Population*, *Intervention*, *Comparison*, *Outcome*, and *Time* (PICO(T)) model to represent different aspects of a research study [28]. Many of the terms are modeled in SNOMED CT. However, the PICO(T) model does not include many of the critical provenance terms that are necessary to reproduce results generated from a scientific study. For example, the PICO(T) model does not represent provenance terms corresponding to the data analysis method (e.g., statistical model used to derive study results), instruments used to record data (e.g., type of blood pressure instrument used in the example research study). To address this issue, we extended the W3C PROV ontology in the ProvCaRe project to model provenance information corresponding to the three aspects of a scientific study, namely Study Tools, Method, and Data.

We extended the PROV-O class `Entity` to model provcare:StudyData, which includes the `provcare:StudyOutcome`, `provcare:ComparisonData`, and `provcare:StudyPopulation` corresponding to the PICO(T) components described earlier. The `provcare` namespace refers to the http://www.case.edu/ProvCaRe/provcare. The class `StudyDesign` represents three categories of biomedical research studies, namely `provcare:FactorialStudy`, `provcare:InterventionalStudy`, and `provcare:ObservationalStudy`. The `FactorialStudy` class is similar to study design class modeled in OBI. The `StudyDesign` class is a subclass of `prov:Plan` class, which also has `StudyConstraint` as a subclass. The `StudyConstraint` represents inclusion and exclusion criteria that are applied to select participants to be recruited into a research study. The ProvCaRe ontology also models multiple classes related to data analysis method as subclass of the `provcare:StudyMethod` class, which is modeled as a subclass of the `prov:Activity` class.

The `provcare:DataAnalysisMethod` class has multiple subclasses, including `provcare:MissingDataProtocol` and `provcare:StatisticalMethod` that model different aspects of data analysis in research studies. The ontology models the two primary categories of statistical analysis methods, namely descriptive analysis and inferential analysis as subclasses of the `StatisticalMethod` class. The ontology also models additional classes representing specific types of statistical analysis techniques as subclasses of the descriptive and inferential analysis classes. The `provcare:StudyInstrument` class is modeled as subclass of `prov:Agent` class and it models instruments used in research studies according to their function and modality. The `StudyInstrument` class includes electrophysiological signal recording instruments (e.g., Electrocardiograph and Electroencephalogram), imaging tools (e.g., MRI), and also software tools (e.g., statistical package R or SAS) as its subclasses. In addition to these ontology classes, the ProvCaRe ontology also extends the PROV-O properties to link ontology classes with appropriate relations. For example, we use OWL2 class-level restriction feature to assert that a `provcare:ResearchStudy` has different types of data collection methods (`provcare:DataCollectionMethod`), such as `provcare:BaselineDataCollectionMethod` representing data collected at start of the study and `provcare:FollowupDataCollectionMethod` representing data collected at subsequent time intervals,

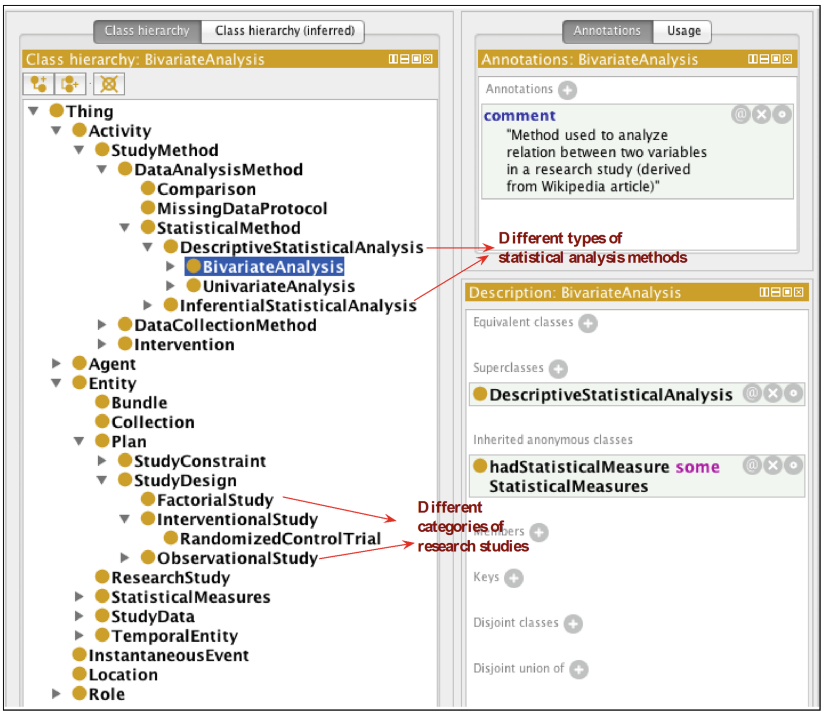

Fig. 2. A screenshot of the ProvCaRe ontology class hierarchy showing different components of the provenance metadata representation framework modeled in the ProvCaRe project

using restriction defined on the object property `provcare:hadDataCollectionMethod` (Fig. 2).

In addition to object properties, the ProvCaRe ontology models additional metadata information about the ontology classes using the RDF(S) annotation properties, for example `rdfs:label`, `rdfs:seeAlso`, and custom properties such as `synonym`. These metadata properties allow provenance applications, such as the ProvCaRe natural language processing (NLP) tool to effectively use the ProvCaRe ontology for entity extraction from biomedical literature. Figure 3 illustrates the class hierarchy of the ProvCaRe ontology. The ProvCaRe ontology provides the required set of precoordinated terms to represent provenance information corresponding to Study Method, Data, and Tools. However, as we discussed earlier we need a well-defined mechanism to create new postcoordinated expressions to represent provenance information and we describe the compositional grammar developed for the ProvCaRe framework in the next section.

3.2 Compositional Grammar Syntax in the ProvCaRe Ontology

We extend and adapt the postcoordination compositional grammar syntax defined to create SNOMED CT expressions to the ProvCaRe framework using classes and

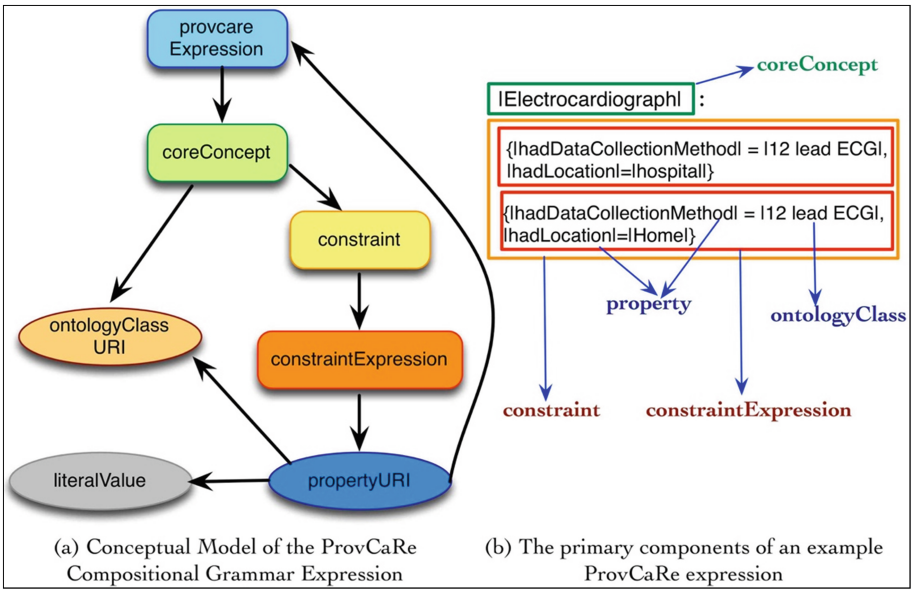

Fig. 3. The conceptual model of the postcoordination expression compositional grammar used in ProvCaRe and an illustrative example are shown.

properties defined in the ProvCaRe ontology and existing biomedical ontologies, for example ontologies listed in NCBO [21]. A ProvCaRe postcoordinated expression consists of a single ontology class, which is the "core" provenance concept of the expression, and a set of properties as well as their values that qualify the core concept. The properties and the associated values may be defined either in the ProvCaRe ontology or NCBO listed ontologies (this ensures that the corresponding ontologies are publicly available) or the values may be RDF literal values (e.g., XML Schema data type). Figure 4 illustrates the conceptual view of the ProvCaRe postcoordination expression syntax. Each ProvCaRe postcoordinated expression is a triple structure with the form "class-property-expression", where the expression is a recursive structure consisting of either a single class or another expression and the "|" symbol is used as start and end delimiters of the terms (similar to the SNOMED CT compositional grammar syntax). The expression refines the core concept with additional values defined for properties and the corresponding concept represented by the expression is a subclass of the core concept.

We use the Augmented Backus-Naur Form (ABNF) [29] to define the ProvCaRe postcoordination expression syntax, which is described in Table 1.

Four categories of postcoordinated expressions can be composed using the compositional grammar in the ProvCaRe framework:

1. **Multi-Concept Expression:** Two or more ontology classes can be combined together using the "+" symbol to form a new concept, which is interpreted to be a

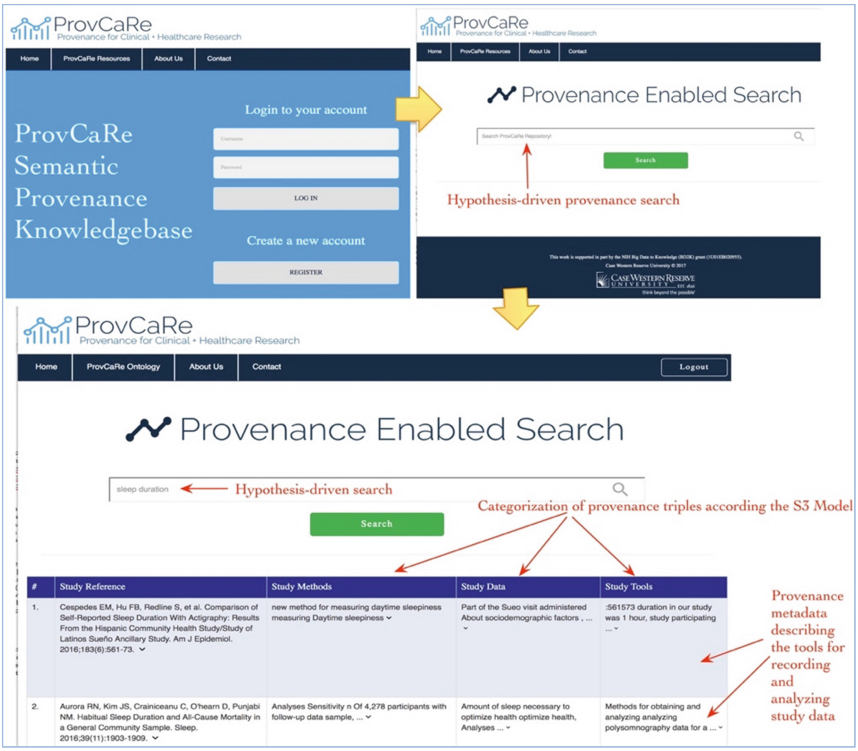

Fig. 4. The user interface of the ProvCaRe semantic provenance knowledgebase available at: https://provcare.case.edu/

subclass of the two original classes. For example, a data analysis method may involve both `provcare:CorrelationAnalysis` and `provcare:CovarianceAnalysis`, which can be represented using the expression[2]: |`CorrelationAnalysis`| + |`CovarianceAnalysis`|.

2. **Concept with Constraints Defined Over Properties:** A provenance term can be refined using additional constraints defined over a ProvCaRe or other ontology property. For example, it is important to record the provenance of blood pressure values of a research study participant in terms of the procedure. This provenance information can be modeled using an expression that combines ProvCaRe ontology and SNOMED CT terms: |`Electrocardiograph`|: |`hadDataCollectionMethod`| = |`12 lead ECG`| (`ECG` has SNOMED CT ID: C0180600 and `12 lead ECG` has SNOMED CT ID: C0430456). A core concept can also include multiple constraints defined using multiple properties. For example, |`Electroencephalogram`|: |`hadStudyInstrument`| = |`Scalp electrode cap`|, |`hadLocation`| = `hospital`| expressions represent the provenance

[2] The namespace for the terms used in the expressions are not repeated for brevity.

Table 1. The specification of the ProvCaRe postcoordination expression syntax with explanatory description.

	Syntax expression	Description
1.	`provcareExpression = subexpression`	A provcare post coordinated expression consists of subExpressions
2.	`subExpression = coreConcept [":" constraint]`	A subExpression consists of a core provenance concept, which is refined through use of constraints, which may consist of multiple constraintExpressions. A subExpression is a subclass of the coreConcept
3.	`coreConcept = "|" ontologyClassURI "|"`	The coreConcept is a provenance ontology class defined in the ProvCaRe ontology
4.	`ontologyClassURI = nonPipe * (* nonPipe)`	The ontology class is listed using the concept ID or namespace aware URI. An ontology class URI may consist of any UTF-8 character except pipe "|" and conform to the URI specification as defined in the W3C URI specifications
5.	`constraint = (constraintExpression) * (["," ] constraintExpression)`	A constraint consists of one or more constraint expressions that are optionally grouped together into a subunit separated by comma
6.	`constraintExpression = ["{"] propertyURI "=" ontologyclassURI/ "("constraintExpression ")"["}"] *constraintExpression`	A constraintExpression consists of an ontology property with an ontology class as value or a constraintExpression as value (for nested expressions) followed by additional constraints
7.	`propertyURI = nonPipe * (*nonPipe)`	The ontology property is listed using the concept ID or namespace aware URI. An ontology property URI may consist of any UTF-8 character except pipe "|" and conform to the URI specification as defined in the W3C URI specifications

information of an EEG in terms of the instrument used to record it and the location of the recording. The two properties `hasStudyInstrument` and `hadLocation` are ProvCaRe and PROV ontology terms respectively.

3. **Concepts with Constraints Defined Over Property Groups:** The ProvCaRe compositional grammar allows grouping of multiple properties into a subunit to reduce ambiguity regarding the ordering of the constraints using an approach that is similar to the SNOMED CT compositional grammar. For example, two ECG recordings may have been conducted at two different locations, which can be represented using the expression |`Electrocardiograph`|: {|`hadDataCollectionMethod`| = |`12 lead ECG`|, |`hadLocation`| = |`hospital`|}, {|`hadDataCollectionMethod`| = |`12 lead ECG`|, |`hadLocation`| = |`Home`|}. The curly braces group two or more properties together to allow humans and software tools to correctly parse the ordering of the constraints in an expression. Similar to the SNOMED CT compositional grammar, the comma between the two subunits is optional.
4. **Concepts with Nested Constraints:** As discussed earlier, the ProvCaRe postcoordinated expressions use a recursive definition, which allows the value in the triple structure of the expressions to be another expression. For example, a research study may use two statistical data analysis techniques that can be represented using the following expression: |`ResearchStudy`|: |`hadDataAnalysisMethod`| = (|`CorrelationAnalysis`| + |`CovarianceAnalysis`|). The nested structure may also be constructed using multiple properties, for example |`ResearchStudy`|: |`hadDataAnalysisMethod`| = (|`StatisticalMethod`|: |`hadStatisticalMeasure`| = |`CentralTendencyMeasure`|).

It is important to note that unlike the SNOMED CT compositional grammar, which allows interpretation of postcoordinated expressions as equivalent or subclass of a given class, the ProvCaRe postcoordinated expressions are interpreted only as subclass of the core concept.

4 Results

We describe the two-fold results of the ProvCaRe project: (1) we demonstrate the practical use of postcoordinated expression in the ProvCaRe ontology using an example research study published by O'Connor et al. [32]; and (2) we demonstrate the effectiveness of the ProvCaRe ontology in the extraction of 38 million provenance triples from 384,802 published articles.

Application of postcoordinated expression: The research study by O'Connor et al. [32] is classified as an `ObservationalStudy` (modeled as subclass of `StudyDesign` in the ProvCaRe ontology). Using the ProvCaRe compositional grammar, we model the data analysis related provenance information using ontology classes and properties modeled in the ProvCaRe ontology and existing biomedical ontologies. For example, the postcoordinated expression |`ResearchStudy`|: |`hadDataAnalysisMethod`| = |`multivariate regression`|, |`hadSoftwareTool`| = |`SAS`|, uses the terms `multivariate regression` and `SAS` modeled in the Bilingual

Ontology of Alzheimer's Disease and Related Diseases (ONTOAD) and Software Ontology (CWO) (listed in NCBO). Similarly, the population selected for the research study can be characterized using the following expression: |ResearchStudy|: {|hadStudyConstraint| = (|StudyExclusionCriterion|: |hadPrescription| = |Antihypertensive medication|)}. This expression represents important provenance metadata describing the constraints used to identify participants for the research study and is essential for other researchers who aim to replicate this study. The method used to collect the data in the research study can also be represented using postcoordinated expression: |ResearchStudy|: {|hadDataCollectionMethod| = |BaselineDataCollection|, |hadDataCollectionMethod| = (|FollowupDataCollection|: |hadTemporalAttribute| = |5 years|)}.

It is important to note that provenance-related postcoordinated expressions need to be created often by domain experts with little or no experience in ontology engineering practices. Therefore, development of a visual user interface form can significantly help domain experts to create valid postcoordinated expression. The ProvCaRe compositional grammar syntax supports the development of a form-based user input template that uses the property values as "widgets" and the corresponding ontology classes as "values". For example, hadDataAnalysisMethod can have a drop-down menu with list of ontology classes corresponding to DataAnalyisMethod or its subclasses. A similar approach is often used in development of ontology-driven user interface applications. The rules defined in the ProvCaRe compositional grammar syntax also supports systematic parsing of the postcoordinated expression, which can be used by provenance applications for validation, querying, and interpretation of research studies annotated with ProvCaRe postcoordinated expressions. The postcoordination-based modeling approach is also extensible as the new provenance-specific terms are modeled in the ProvCaRe ontology, for example detailed representation of how missing data is handled in research studies, and new biomedical ontologies are released through NCBO. This is an important feature of the proposed approach as the ProvCaRe project extracts and analyses provenance metadata information from additional biomedical domains, such as neurological disorders and lung cancer, as part of our ongoing and future work.

Creation of the ProvCaRe Semantic Provenance Knowledgebase: The extraction of structured data from free text is a significant challenge and this has been a focus of extensive research in computer science using statistical machine learning as well as rule-based techniques [30]. The use of Semantic Web techniques especially using ontologies as reference knowledge model has been an effective approach for natural language processing (NLP) [31]. However, we are not aware of any previous work that use ontologies for extracting provenance metadata from unstructured text. To extract and analyze the provenance information from published biomedical research studies, we have developed a novel Natural Language Processing (NLP) workflow using the ProvCaRe ontology [9]. Using the ontology-enabled NLP workflow, we have processed and extracted provenance information from 384,802 published articles describing biomedical research studies (the articles are available from the National Center for Biomedical Informatics PubMed resource, https://www.ncbi.nlm.nih.gov/pubmed/). We extracted more than 38 million provenance triples from these published articles by

using the ProvCaRe ontology for named entity recognition (NER) and predicate identification. These provenance triples are available for querying and analysis in the ProvCaRe semantic provenance knowledgebase, which can be accessed at: https://provcare.case.edu/ (Fig. 4 shows the details of the ProvCaRe knowledgebase).

As far as we know, the ProvCaRe knowledgebase with 38 million provenance triples is one of the largest real world dataset of biomedical provenance information available to the research community for querying and analysis. The knowledgebase supports querying using two approaches: (1) users can use a "hypothesis-driven" query approach to search for previous research studies corresponding to a given hypothesis and view the provenance metadata associated with each of these studies for scientific reproducibility; and (2) use the provenance information of previous studies to design new experiments with rigorous protocols for ensuring transparent reporting as well as supporting reproducibility. As shown in Fig. 5, the provenance triples extracted from the published articles are classified into one of three categories of provenance metadata defined in the ProvCaRe S3 model. Table 2 lists the distribution of provenance triples in each of the three categories.

The distribution of provenance triples in Table 2 demonstrates that provenance

Table 2. The number and distribution of provenance triples in the ProvCaRe knowledgebase according to the S3 model

	Distribution of Provenance Triples (total: 38.47 million provenance triples)		
	Study methods	Study data	Study instruments
Total number of triples	12,212,129	15,361,311	10,905,018
Percent distribution of triples	32%	40%	28%

metadata describing the method and data of research experiments is well-described in published articles, however there is limited provenance metadata describing the instruments used in research studies. This highlights an important limitation of published articles describing research studies as the instruments used in an experiment and the parameters used to record experiment data are essential for reproducibility of scientific results. We believe new guidelines and best practices, for example the NIH Rigor and Reproducibility guidelines can help address these issues in transparent reporting of new research experiments.

5 Conclusions and Future Work

Our work was motivated by the need to represent provenance metadata information describing research studies in a variety of biomedical domains for scientific reproducibility. With the known limitations of modeling large number of classes and properties in a single ontology using precoordinated modeling approach, we extended and adapted the SNOMED CT compositional grammar syntax to create ProvCaRe postcoordinated expressions. The ProvCaRe postcoordinated expressions use

provenance-specific classes and properties defined in the ProvCaRe ontology and re-uses terms from existing biomedical ontologies to represent provenance metadata. The ProvCaRe ontology extends the W3C PROV ontology to represent three core provenance terms: Study Method, Data, and Tools. We define the ProvCaRe compositional grammar syntax using ABNF notation and define four categories of postcoordinated expressions that can be created to represent provenance information. We demonstrate the application of the ProvCaRe postcoordinated expressions in modeling the provenance information associated with a research study and the use of the ProvCaRe ontology in the creation of one of the largest semantic provenance knowledgebase with more than 38 million provenance triples.

Acknowledgement. This work is supported in part by the National Institutes of Biomedical Imaging and Bioengineering (NIBIB) Big Data to Knowledge (BD2K) grant (1U01EB020955) NSF grant# 1636850

References

1. Collins, F.S., Tabak, L.A.: Policy: NIH plans to enhance reproducibility. Nature **505**, 612–613 (2014)
2. Landis, S.C., Amara, S.G., Asadullah, K., et al.: A call for transparent reporting to optimize the predictive value of preclinical research. Nature **490**(7419), 187–191 (2012)
3. Redline, S., Dean III, D., Sanders, M.H.: Entering the era of "Big Data": getting our metrics right. SLEEP **36**(4), 465–469 (2013)
4. Baker, M.: 1,500 scientists lift the lid on reproducibility. Nature **533**(7604), 452–454 (2016)
5. NIH: Principles and Guidelines for Reporting Preclinical Research (2016). https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research. Accessed 20 July 2017
6. Buneman, P., Davidson, S.: Data provenance - the foundation of data quality (2010)
7. Goble, C.: Position statement: musings on provenance, workflow and (semantic web) annotations for bioinformatics. In: Workshop on Data Derivation and Provenance, Chicago (2002)
8. Sahoo, S.S., Sheth, A., Henson, C.: Semantic provenance for escience: managing the deluge of scientific data. IEEE Internet Comput. **12**(4), 46–54 (2008)
9. Valdez, J., Kim, M., Rueschman, M., Socrates, V., Redline, S., Sahoo, S.S.: ProvCaRe semantic provenance knowledgebase: evaluating scientific reproducibility of research studies. Presented at the American Medical Informatics Association (AMIA) Annual Conference, Washington DC (2017)
10. Zhao, J., Goble, C., Stevens, R., Turi, D.: Mining Taverna's semantic web of provenance. J. Concurr. Comput. Practice Exp. **20**(5), 463–472 (2008)
11. Simmhan, Y.L., Plale, A.B., Gannon, A.D.: A survey of data provenance in e-science. SIGMOD Rec. **34**(3), 31–36 (2005)
12. Moreau, L., Clifford, B., Freire, J., et al.: The open provenance model core specification (v1.1). Future Gener. Comput. Syst. **27**(6), 743–756 (2010)
13. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S.: OWL 2 Web Ontology Language Primer. In: W3C Recommendation. World Wide Web Consortium W3C (2009)

14. Sahoo, S.S., Sheth, A.: Provenir ontology: towards a framework for eScience provenance management. Presented at the Microsoft eScience Workshop, Pittsburgh, USA, October 2009
15. Moreau, L., Missier, P.: PROV Data Model (PROV-DM). In: W3C Recommendation. World Wide Web Consortium W3C (2013)
16. Lebo, T., Sahoo, S.S., McGuinness, D.; PROV-O: the PROV ontology. In: W3C Recommendation. World Wide Web Consortium W3C (2013)
17. Cheney, J., Missier, P., Moreau, L.: Constraints of the PROV data model. In: W3C Recommendation. World Wide Web Consortium W3C (2013)
18. Dean, D.A., Goldberger, A.L., Mueller, R., Kim, M., et al.: Scaling up scientific discovery in sleep medicine: the National Sleep Research Resource. SLEEP **39**(5), 1151–1164 (2016)
19. Cyganiak, R., Wood, D., Lanthaler, M.: RDF 1.1 concepts and abstract syntax. In: W3C Recommendation, World Wide Web Consortium (W3C) (2014)
20. Rector, A., Luigi, I.: Lexically suggest, logically define: quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT. J. Biomed. Inform. **45** (2), 199–209 (2012)
21. Musen, M.A., Noy, N.F., Shah, N.H., Whetzel, P.L., Chute, C.G., Story, M.A., Smith, B.: NCBO team: The national center for biomedical ontology. J. Am. Med. Inform. Assoc. **19** (2), 190–195 (2012)
22. Köhler, S., Doelken, S.C., Mungall, C.J., et al.: The human phenotype ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. **42**, 966–974 (2014). Database Issue
23. Giannangelo, K., Fenton, S.: SNOMED CT survey: an assessment of implementation in EMR/EHR applications. Perspect Health Inf. Manag. **5**, 7 (2008)
24. Bodenreider, O., Stevens, R.: Bio-ontologies: current trends and future directions. Brief. Bioinform. **7**(3), 256–274 (2006)
25. Sim, I., Tu, S.W., Carini, S., Lehmann, H.P., Pollock, B.H., Peleg, M., Wittkowski, K.M.: The ontology of clinical research (OCRe): an informatics foundation for the science of clinical research. J. Biomed. Inform. **52**, 78–91 (2014)
26. Tu, S.W., Peleg, M., Carini, S., Bobak, M., Ross, J., Rubin, D., Sim, I.: A practical method for transforming free-text eligibility criteria into computable criteria. J. Biomed. Inform. **44** (2), 239–250 (2011)
27. Bandrowski, A., Brinkman, R., Brochhausen, M., et al.: The ontology for biomedical investigations. Plos One **11**(4), e0154556 (2016)
28. Huang, X., Lin, J., Demner-Fushman, D.: Evaluation of PICO as a knowledge representation for clinical questions. Presented at the AMIA Annual Symposium Proceedings (2006)
29. Overell, P.: Augmented BNF for Syntax Specifications: ABNF. https://tools.ietf.org/html/rfc5234. Accessed 20 Aug 2017
30. Hearst, M.A.: Untangling text data mining. In: 37th the Association for Computational Linguistics on Computational Linguistics meeting, pp. 3–10 (1999)
31. Rindflesch, T.C., Pakhomov, S.V., Fiszman, M., Kilicoglu, H., Sanchez, V.R.: Medical facts to support inferencing in natural language processing. Presented at the AMIA Annual Symposium Proceedings (2005)
32. O'Connor, G.T., Caffo, B., Newman, A.B., Quan, S.F., Rapoport, D.M., Redline, S., Resnick, H.E., Samet, J., Shahar, E.: Prospective study of sleep-disordered breathing and hypertension: the sleep heart health study. Am. J. Respir. Crit. Care Med. **179**(12), 1159–1164 (2009)