Provenance Management in BioSciences[1]

Sudha Ram, Jun Liu

430J McClelland Hall, Department of MIS, Eller School of Management,
University of Arizona, Tucson, AZ, 85721,
ram@eller.arizona.edu, jliu@email.arizona.edu

Abstract: Data provenance is becoming increasingly important for biosciences with the advent of large-scale collaborative environments such as the *iPlant* collaborative, where scientists collaborate by using data that they themselves did not generate. To facilitate the widespread use and sharing of provenance, ontologies of provenance need to be developed to enable the capture and standardized representation of provenance for biosciences. Working with researchers from the iPlant Tree of Life (iPToL) Grand Challenge Project, we developed a domain ontology of provenance for phylogenetic analysis. Relying on the conceptual graph formalism, we describe the process of developing the provenance ontology based on the W7 model, a generic ontology of data provenance. This domain ontology provides a structured model for harvesting, storing and querying provenance. We also illustrate how the harvested data provenance based on our ontology can be used for different purposes.

Keywords: Provenance, tree of life, W7 model, conceptual graphs.

1 Introduction

In recent years, the tendency toward "big science" (i.e., large-scale collaborative science) is increasingly evident in the biological sciences - facilitated by a breakdown of the traditional barriers between academic disciplines and the application of technologies across these disciplines. The growing number and size of computational and data resources is enabling scientists to perform advanced scientific tasks in large collaborative scientific projects such as the the *iPlant* Collaborative (iPlant, http://www.iplantcollaborative.org). Provenance is becoming increasingly important for biosciences as more scientists collaborate by using data that they themselves did not generate. Tracking data provenance helps ensure that data provided by many different providers and sources can be trusted and used appropriately. Data provenance also has several other critical uses, including data quality assessment, generating data replication recipes, data security management, and others as outlined in [1].

Recently, a consensus has emerged on the need to develop a generic ontology for standardized, application- and organization-independent representation of data provenance [2]. Such a generic ontology will allow provenance to be exchanged between systems. More importantly, a generic ontology is meant to be extensible and shared across applications and modified according to the requirements of a particular domain, thus eliminating the need to develop domain ontologies from the very beginning. Based on analyzing over 100 use cases, we developed a generic ontology

[1]This research is supported in part by research grants from the National Science Foundation Plant Cyberinfrastructure Program (#EF-0735191) and from the Science Foundation of Arizona.

of provenance called the W7 model [3, 4] that defines provenance as consisting of seven interconnected components including what, how, who, when, where, which and why. The W7 model was designed to be general and comprehensive enough to cover a broad range of provenance-related vocabularies (i.e., concepts and their relations). However, the W7 model alone, no matter how comprehensive, is insufficient for capturing provenance for all types of data in biosciences without being adapted and extended. The types and level of detail for tracking provenance vary by data type, purpose, discipline, and project. For instance, the provenance of data on a plant gene may include not only the experimental process by which it was derived, but also information about what plant part and sample was used and how the sample was manipulated. The objective of this paper is to illustrate the process of developing a domain ontology for the plant science domain by adapting and extending the W7 model. Our work is set within the context of the iPlant collaborative project (www.iplantcollaborative.org). The purpose of iPlant is to develop a cyberinfrastructure that enables the plant sciences community to collaboratively define, investigate and solve the grand challenges of plant biology.

2 Background

2.1 The iPlant Collaborative

The iPlant Collaborative (iPlant) project's mission is to foster the development of a diverse, multidisciplinary community of scientists, teachers, and students, and a cyberinfrastructure that facilitates significant advances in the understanding of plant science through the application of computational thinking and approaches to Grand Challenge problems in plant biology. The plant sciences community has identified two important grand challenges they need to address. The first grand challenge is called iPlant tree of Life (iPTOL) while the second one is called iPlant Genotype to Phenotype (iPG2P). Our focus in this paper is on the development of a provenance tracking and management mechanism for the iPTOL grand challenge.

Knowledge of evolutionary relationships is fundamental to biology, yielding new insights across the plant sciences, from comparative genomics and molecular evolution, to plant development, to the study of adaptation, speciation, community assembly, and ecosystem functioning. Although our understanding of the phylogeny of the half million known species of green plants has expanded dramatically over the past two decades, the task of assembling a comprehensive "tree of life" for them presents a Grand Challenge. Its solution will require a significant intellectual investment at the developing intersection between phylogenetic biology and the computer sciences. iPTOL brings together plant biologists and computer scientists to build the cyberinfrastructure needed to scale up phylogenetic methods by 100-fold or more, to enable the dissemination of data associated with such large trees, and to implement scalable "post-tree" analysis tools to foster integration of the plant tree of life with the rest of the botanical sciences. The undertaking to unravel the evolutionary relationships among all living things, and to express this in the form of a phylogenetic tree of life, is one of the most profound scientific challenges ever undertaken, and represents a true "moonshot" for plant sciences. We anticipate that early success in addressing the plant phylogeny problem will be especially useful in connection with other Grand Challenge Projects supported through the iPlant Collaborative that involve comparisons between genes, genomes, or species, insuring a broad impact of the project as a whole. Finally, the plant tree of life provides exciting opportunities for training and outreach at all levels. Since Darwin, the tree of life has proven to be a very accessible visual metaphor for nonscientists, providing an elegant opening for communicating results in the plant sciences and evolutionary biology to people with diverse backgrounds.

Data provenance is critical for iPTOL. It serves three major purposes: 1) to evaluate the quality and trustworthiness of data, 2) to determine how data has been processed and modified data within the discovery environment in iPlant, and 3) to enable proper attribution of the creator/owners of the datasets and the researchers' discoveries. In this paper, we describe the development of a domain ontology of provenance for iPToL by extending the W7 model [3, 4] . Extending a generic ontology such as the W7 model to accommodate domain specific requirements can be challenging for domain experts unless a structured approach is followed. We describe the procedure for extending the W7 model, which can be applied by other domain experts who intend to adopt and extend the W7 model for their own fields.

2.2 The Generic W7 model for Data Provenance

Based on analysis a large number of use cases collected from various domains, we conceptualized provenance as a set of 7-tuple, (*what, when, where, how, who, which, why*), and developed a generic ontology of provenance called the W7 model.

The anchor of our provenance is *what*, i.e., events that affect a data object during its lifecycle. An event can be content related (e.g., *creation* and *modification*) or non content related (e.g., *location change*, *ownership change*, *format change*, *right change, access* and *annotation event*). Provenance of a data object includes events ranging from creation, to its modification, to its final destruction and archiving. The relationships between *what* and the other six Ws are graphically represented in Fig. 1.The other six Ws including *when, where, how, who, which,* and *why* are linked to *what* associated with a data object. The further classification of *what* and the other w's is shown in Fig. 2.

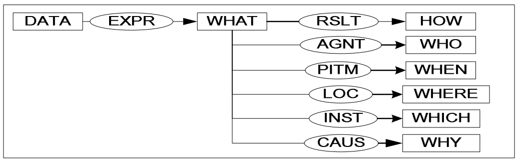

Fig. 1. Relationship between *what* and the other w's

How represents an *action* leading to the event. It can be classified into *single action* and *complex action*. For instance, *purchase* and *donation* are actions that lead to an ownership change. *When* represents the *time* of the event. *Where*, by default, represents the *location* of the event. An event such as a location change is associated with two locations: origin and destination. *Origin*, i.e. where the data came from, is critical provenance information and thus captured as a subtype of *where*. It is common for a digital record to travel from system *a* to system *b* while retaining its original copy in *a*.

Such an event is considered a data creation, and origin is important *where*-provenance for the event. *Who* represents *people* or *organizations* involved in the event. It includes *agents* who initiated the event as well as *participants* of the event. *Why* refers to *reasons* that explain why an event occurred. In our research, *why* includes *belief* and *goal*. A belief refers to the *rationale* or *assumptions* made in generating or modifying the data. Our use cases indicate that a common goal in creating or manipulating data is to use it in a *project* or an *experiment*. Finally, *which* refers to *instruments* or *software programs* used in the event.

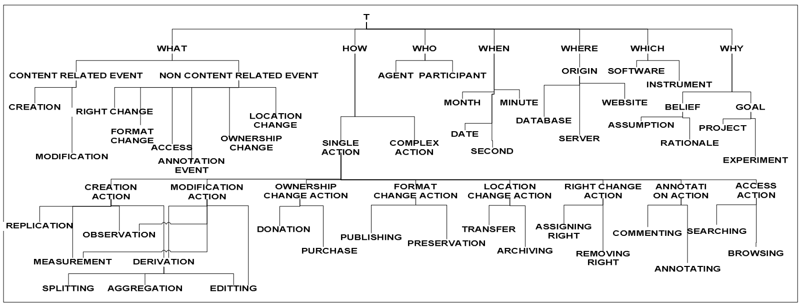

Fig. 2. Hierarchy of the 7 w's (⊤ represents the universal type, a super type all other types)

We represent the W7 model using the conceptual graph (CG) formulism developed by Sowa[5]. We briefly introduce the basic conceptual graph formalism.

1) Conceptual graphs (CGs)

A conceptual graph is a finite, connected, bipartite graph with nodes of one type called *concepts* and nodes of another type called *conceptual relations*. The conceptual graph shown in Fig. 3 conveys the proposition that "the creation of the data #115 was made by Nicole". The boxes are concepts. A concept is made up of either a concept type alone or a concept type and its referent information. In the example, the concept [creation] is *generic* with only a type label inside the box. The other concepts are *individual*. They have a colon after the type label, followed by a *name* (e.g., Nicole) or a unique identifier called an *individual marker* (#115), representing a specific instance of the type. The ovals are conceptual relations. The conceptual relations labeled OBJ and AGNT shown in Fig. 3 represent the linguistic cases *object* and *agent* of case grammar.

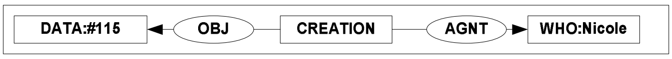

Fig. 3. A conceptual graph example

To distinguish the graphs that are meaningful in a domain of interest from those that are not, certain graphs are declared to be *canonical*. The CG model represents the knowledge in a domain of interest using two components: a *canon* and a set of *conceptual graphs* that are canonical. The canon contains the information necessary for deriving the conceptual graphs. It has four components: a type hierarchy *T*, a set of individual markers *I*, a conformity relation :: that relates type labels in *T* to markers in *I*, and a finite set of canonical graphs *B*, called the *canonical basis*. In essence, the canon provides a repository of concepts and relations to build conceptual graphs. Not all assemblies of concepts and relations into a conceptual graph are meaningful or “canonical”. The canon provides a finite set of canonical graphs that indicate a permissible combination of concepts and relations as the canonical basis. A large number of conceptual graphs that are canonical can then be derived from those in the canonical basis by application of the canonical formation operations. Each of them is a representation of a part of knowledge under the canon. It could thus be considered that conceptual graphs represent knowledge itself while the canon acts as a framework for the organization of knowledge and helps encourage a disciplined approach to representing knowledge in the CGs.

2) The W7 model represented in the CG formulism

Our generic ontology is called the W7 model since we conceptualize provenance as a sequence of seven w’s including *what, when, where, how, who, which,* and *why*. Based on the CG formulism, we define the W7 model as a triple $W7 = (T_c, S, W7Graph)$ whose components are defined below.

T_c is a concept type hierarchy. It includes provenance-related concept types organized in a hierarchical structure, as shown in Fig. 2. S represents a set of schemas defined for concepts located in T_c. A schema is a structure of knowledge that corresponds to a particular concept type t in T_c. Formally, a schema in CGs is defined as a monadic abstraction λau where the formal parameter a is of type t for which the schema is defined, and the body u is a conceptual graph that provides the background of what is plausibly true about the concept type t. The CG formulism allows us to attach any number of “related” schemas to a concept. Schema definition is thus a critical mechanism for ontology extension. For the purpose of the current research, we partition S into two sets: optional schemas and mandatory schemas. A schema of a concept is by default optional that state the commonly associated properties. A schema may not be true or necessary for every use of the type. A mandatory schema, on the other hand, defines necessary conditions that include mandatory properties of the concept. Fig. 4 shows several schema schemas that belong to S. The conceptual graph shown in Fig. 4(a) represents a mandatory schema for the concept type DERIVATION. It asserts that a derivation must have some input data. The CG shown in Fig. 4(b) is an optional schema for the concept type SINGLE ACTION, representing one way the concept can be used: a single action has another action as its successor.

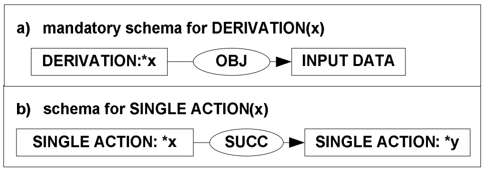
a) mandatory schema for DERIVATION(x)
DERIVATION:*x
OBJ
INPUT DATA
b) schema for SINGLE ACTION(x)
SINGLE ACTION: *x
SUCC
SINGLE ACTION: *y

Fig. 4. Example schemas

W7Graph, or the W7 graph, is the graph represented in Fig. 1 that includes the seven W’s and indicates the relationship between them. The W7 graph serves as a base graph, representing the overall structure of provenance. When representing provenance for a given type of data, this graph must be specialized. Several types of important graphical operations called the canonical formation rules (including *copy*, *simplify*, *restrict* and *join*) allow a number of more specialized conceptual graphs to be derived from the base graph. The *copy* rule builds an exact copy of a given graph or its subgraph. The *simplify* rule removes duplicate relations in a graph. One can *restrict* a concept by replacing the label of that concept type with a subtype (e.g., WHAT can be restricted to CREATION). *Restrict* can also replace a generic concept with an individual instance. The *join* rule merges identical concepts. Concepts are identical if both the concept type and referent are the same. The merge is achieved by overlaying one graph on the other at the point that they are identical. Fig. 5 illustrates some of formation rules. Suppose we want to derive a graph that asserts "some data is created through a derivation performed upon some input data" (i.e., the graph e in Fig. 5). This graph can be derived from the W7 graph shown in Fig. 1. The graph a is a copy of a subgraph of the W7 graph shown in Fig. 1. The graph b in Fig. 5 results from restricting the type WHAT in the graph a) to CREATION, and the graph c is the result of restricting HOW in the graph b to DERIVATION. The graph d represents a schema defined for the concept DERIVATION (see Fig. 4). Then join can merge the two concepts of type DERIVATION in the graph c and d to form the graph e.

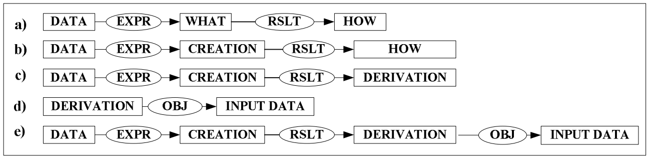

Fig. 5. Formation rules

3 Developing a Domain Ontology for the iPToL Project

Developing a domain ontology for the iPToL project by extending the W7 model involves several steps. First, we identify different types of data available in the domain. The primary type of data available in iPToL is on phylogenetic trees. There are other types of data such as trait data of the species in the phylogenetic trees and outputs from different analysis activities such as phylogenetically independent contrast analysis (*PIC*). Second, for a given type of data such as trees, we identify the different types of events that may affect the data over its lifetime. For instance, a tree file can be created and modified. There are also annotation events in which as images can be added to specific nodes of a tree. In iPToL, researchers can also share a tree file with other people by assigning access rights to other people. Third, we determine the *how, who, when, where, which,* and *why* associated with each type of event that can affect a type of data or data object. Let's consider *how* first. *How* refers to actions leading up to an event. In iPTol, researchers can perform several different types of actions to create a new tree. They can import/upload a tree file from an existing source such as TreeBase or MorphoBank, edit an existing tree and then save it as a new one, or create a tree by merging existing trees. They can also perform different editing actions to modify an existing tree, such as change the name or branch length of a species node in the tree, change the layout of the tree, or add or delete one or more species nodes. They can also first reconcile a tree file and its trait data and then, as a result of the reconciliation, remove unmatched species or swap some species nodes. *Who* then represents the agent performing the event. *When* records the time of the event. *Where*, i.e., where the data came from, is critical for data that were imported from external sources. *Which* represents software (such as Phylowidget or Phylomatic) used to modify or merge existing trees. In iPToL, if a tree file was modified, researchers need to *why* the modification occurred. After defining the events (*what*) relevant to each type of data or data object, and the *how, who, when, where, which,* and *why* for each event, we construct a domain ontology for the iPToL project.

A domain ontology for iPToL consists of a set of 5-tuple (T_o, S_o, *W7Graph*, *E*). We define such a 5-tuple for each type of data in iPToL. Here, we present a specific one defined for phylogenetic trees, and describe each of its components.

T_o is a concept type hierarchy. We developed it based on T_c in the W7 model by adding some domain specific concepts and then pruning the ones that are not applicable or relevant for iPToL. Fig. 6 represents the type hierarchy defined for phylogenetic trees. Compared with the type hierarchy of the generic ontology shown in Fig. 1, it includes a number of domain specific concepts. For instance, different types of tree editing actions such as *add species*, *delete species, edit species*, and *change layout* were included in it. *Edit species* were further classified into *change name* and *change branch length*.

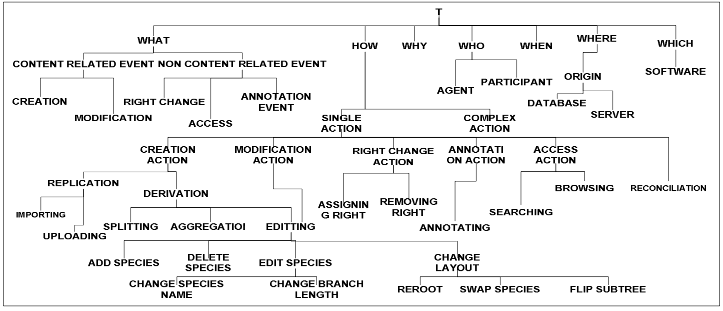

Fig. 6. Concept type hierarchy in iPToL ontology

S_o represents a set of schemas defined for concept types in T_o. If a concept type in T_c was retained in T_o, then the schemas defined for the concept in the W7 model would be imported to the domain ontology. It may include schemas defined for the newly added concept types, and we may also define additional schemas for the retained ones, specifying a new way the concept type can be used. These schemas would be used to provide background about a concept. We define a mandatory schema shown in Fig. 7(a) for the concept type *reconciliation*, indicating that a reconciliation must be performed on a phylogenetic tree and some trait data. The schema shown in Fig. 7(b) specifies that a *replication* must have some source data and the source data has its own provenance. Fig. 7(c) represents a schema defined for *software.* This is optional, which means we can choose to record the author and version for the software used in performing the action.

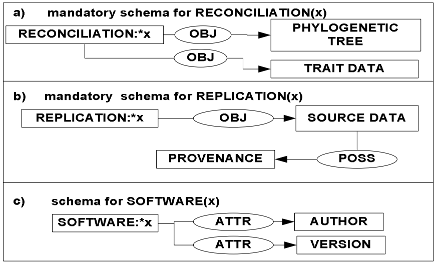

Fig. 7. Schemas defined in the iPToL ontology

W7Graph remains the same as the one shown in Fig. 1. In the CG vocabulary, these three components - T_o, S_o, and *W7Graph* – form a *canon*, i.e., information necessary for deriving other canonical conceptual graphs. We derive a set of canonical graphs based on the canon. Each of them represents a combination of what, how, who, when, where, which, and why, that is meaningful and relevant for representing provenance for Phylogenetic trees. These canonical graphs are termed *event graphs*, since each of them describes the information associated with one type of event that can affect a type of data or data object, in our example this is phylogenetic trees.

E represents a set of *event graph*s. A number of event graphs have been created. Due to space limitations, we show two examples of an event graph in Fig. 8. The event graph shown in Fig. 8(a) specifies that the creation (what) of a phylogenetic tree can be a result of importing (how) some source data with its own provenance from a database (where) by an agent (who) at certain time (when) for some reason(why).*Which* was not included in this graph since it was deemed of little use for this type of event. This graph is derived from the W7 graph shown in Fig. 1 by performing a sequence of graphical operations described earlier. These operations include a *copy* of a subgraph (without *which*) of the W7 graph, several *restrict* operations (e.g., *what* is restricted to *creation*, *how* to *importing*, and *where* to *database*) according to the type hierarchy show in Fig. 6 and then a join with the mandatory schema shown in Fig. 7(b), which was defined for *replication* and thus applicable to *importing*, a subtype of *replication*. The event graph shown in Fig. 8(b) describes how a phylogenetic tree can be modified. A tree can be *modified* as a complex action including *reconciliation* and then *editing*. The reconciliation is performed using the tree and some trait data. We also attempt to capture the *who*, *when*, *why* associated with the *modification* event as well as which *software* was used and the *author* and *version* of the software. Similarly, this graph was also derived from the W7 graph based on a sequence of graphical operations. Relying on different mechanisms provided by the CG formalism, we construct event graphs that represent the domain specific provenance in a *disciplined* yet *flexible* way. Our approach is *disciplined* since the event graphs can only be canonically defined from the W7 model that defined the structure of provenance. Our approach is also *flexible* since the schema definition and the join operation enable us to attach schemas to a concept to provide background about the concept at any level of detail. The event graphs are used as conceptual schemas for capturing the provenance. Provenance captured for an event that affects a tree is an instantiation of one of the event graphs.

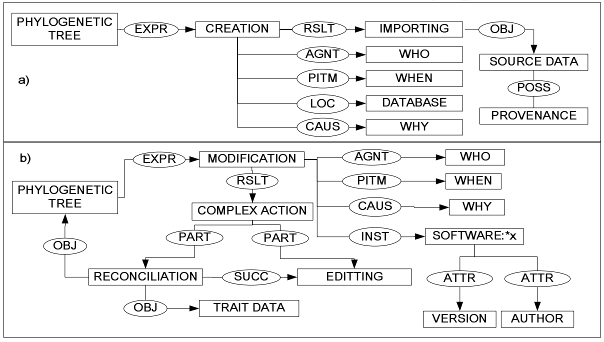

Fig. 8. Examples of event graphs

4 Using Provenance

As discussed previously, provenance serves three major purposes for iPToL: 1) *Data Quality Assessment*: It helps evaluate the quality and trustworthiness of data, 2) *Replication Recipes*: It allows plant scientists to understand how data was processed and modified within a discovery environment in iPlant, and 3) *Attribution:* It enables proper attribution of the creator/owners of the datasets and the researchers' discoveries. Different provenance information may contribute to different uses of data provenance.

A variety of provenance information can be used for estimating the quality of the data. For instance, *where* the data came from is critical for understanding the quality of data. After a tree file is imported, *who* modified it for what purposes (why) is of utmost importance to determine data quality. We have developed a mechanism in the iPlant discovery environment that enables users to visually browse the whole provenance of any dataset or data object to understand and evaluate its data quality. For instance, the provenance of a tree file "PDAP.tree.nex" is shown in Fig. 9. The file was imported by a person named Nicole from TreeBASE. It was then modified by Doug. He changed the name of a species to be consistent with a naming convention used elsewhere. This tree file was then modified by Nicole. She reconciled the tree file with its trait data, and subsequently removed a species in the tree file.

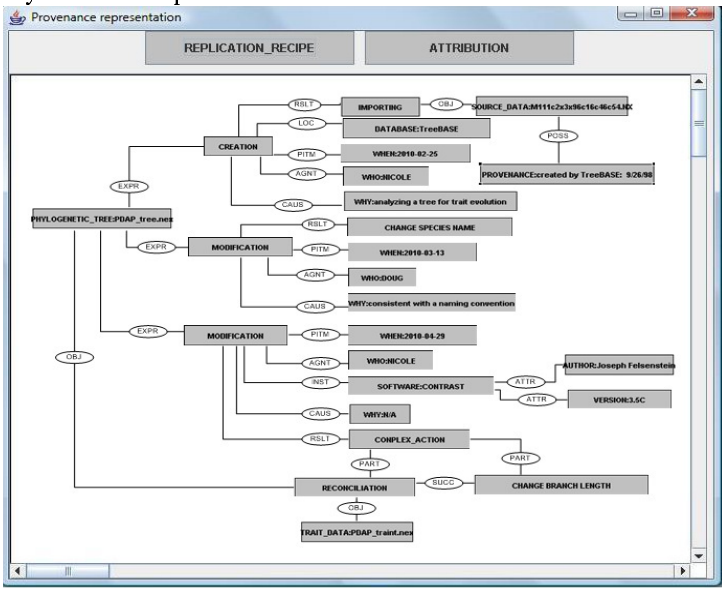

Fig. 9. Graphical representation of provenance

Another use of provenance is to provide a replication recipe for data. Fig. 10 shows a scientist who browses the provenance to understand *how* the data was processed and *which* software tool was used to manipulate it.We also have mechanisms to query and browse the *who* provenance since attribution of the creator/owners of the datasets and the researchers' discoveries, on the other hand, relies primarily on provenance such as who created and modified the data.

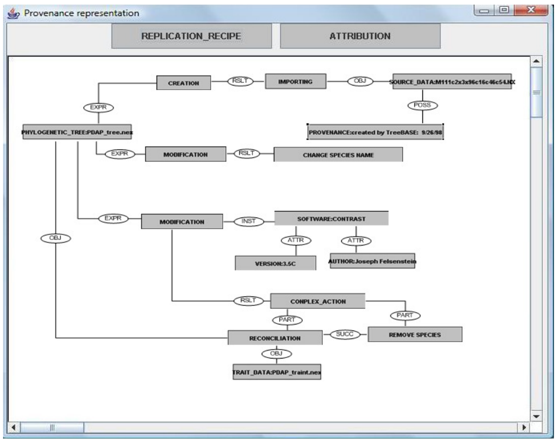

Fig. 10. The *how* and *which* associated with the events

5. Conclusion and Future Research

In this paper, we described how a domain ontology of provenance was developed for the iPlant Tree of Life (iPToL) Grand Challenge Project by extending a generic ontology of provenance in the form of the W7 model. Our approach for developing a domain ontology of provenance can be applied by other domain experts who intend to adopt and extend the W7 model for their own fields.

In our future work, we propose to investigate the uses of provenance. Representing provenance in a structured way enables more sophisticated uses of provenance. As an example, we are developing metrics mapped to different components of provenance (e.g., the *how* or *wh*o provenance) that can be used to assessing the quality of data semi-automatically. We are also extending this work to other grand challenges in iPlant as well as other domains. This is crucial in most bioscience applications, since an ontology that enables the capture and standardized representation of provenance are critical for scientific data sharing.

References

1. Y. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance Techniques," Indiana University, Technical Report IUB-CS-TR618, 2005.
2. L. Moreau, J. Freire, J. Futrelle, R. E. McGrath, J. Myers, and P. Paulson, "The Open Provenance Model: An Overview," *Lecture Notes in Computer Science*, vol. 5272, pp. 323-326, 2008.
3. S. Ram and J. Liu, "Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling," *Lecture Notes in Computer Science 4512*, pp. 17-29, 2007.
4. S. Ram and J. Liu, "A New Perspective on Semantics of Data Provenance," presented at the First International Workshop on the Role of Semantic Web in Provenance Management, Washington D.C., 2009.
5. J. Sowa, *Conceptual structures: Information processing in Mind and Machine*. Reading, MA: Addison-Wesley, 1984.