Using Domain Ontologies to Help Track Data Provenance

Renato Fileto[1,3] Claudia Bauzer Medeiros[1] Ling Liu[2] Calton Pu[2] Eduardo Delgado Assad[3]

[1]Institute of Computing, University of Campinas
Caixa Postal 6176, Campinas, SP, 13081-970 - BRAZIL
{fileto|cmbm}@ic.unicamp.br

[2]College of Computing, Georgia Institute of Technology
801 Atlantic Drive, Atlanta, GA, 30332-0280 - USA
{lingliu|calton}@cc.gatech.edu

[3]Embrapa Information Technology, Brazilian Agricultural Research Corporation
Av. Dr. Andre Torsello, 209, Campinas, SP, 13083-886 - BRAZIL
{assad|fileto}@cnptia.embrapa.br

Abstract

Traditional techniques for tracking data provenance have difficulty adapting to the dynamics of the Web. This paper proposes a scheme for provenance estimation, based on domain ontologies. This scheme is part of the POESIA approach for multi-step integration of semi-structured data. The ontologies used for tracking provenance also help to describe, discover, reuse and integrate data and services. In contrast to traditional techniques, this scheme derives data provenance with fewer annotations at the extensional level and thus lower maintenance costs. Additionally, it promotes the use of ontologies to categorize and correlate scopes of data sets, thereby capturing the operational semantics of data integration processes.

1. Introduction

Data provenance (also called data *genealogy* or *pedigree*) is the description of the origins of a piece of data and the process by which it was produced [1]. Consider information systems involving data integration, such as data warehouses or the dissemination, collection and processing of scientific data over the Web (e.g., climate or biological data). It is imperative to track the gathering, processing and propagation of data across potentially distributed and autonomous systems [2]. The knowledge about the data sources and subsequent processing history of a piece of data is fundamental to assess the quality and usefulness of this data for a particular application. Data provenance is also necessary to investigate the cause of erroneous values detected at any point of the processing flow.

The problem of determining the data provenance of a data item or data set has been studied in a variety of settings, e.g., for data warehouses using relational data sources or in the context of cooperative processes with data exchange in several formats. The solutions proposed in the literature usually involve some kind of annotation or the "inversion" of the functions/queries used to transform data.

The Internet poses new challenges for provenance tracking. The autonomy of the components and the multi-institutional nature of Web applications results in a profusion of data contents, demanding self-describing data sets. Traditional approaches for tracking data provenance, relying on detailed descriptions and tight control of the data transformation flow, cannot be easily adapted to the Web. Detailed information about distributed data processing on the Web, such as the queries/functions used to transform and move data across sites, are often unavailable. A better solution in this context is to build a general framework for provenance tracking, including when necessary detailed analysis of specific portions and emphasizing the semantics of data and processes.

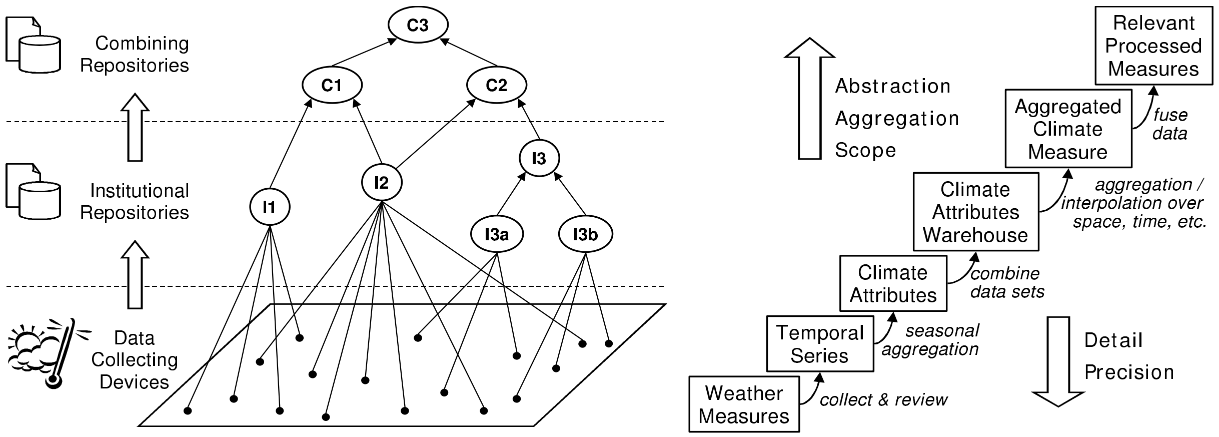
POESIA [3, 4] (Processes for Open-Ended Systems for Information Analysis) is an approach for multi-step integration of semi-structured data in an open and distributed environment. Inspired by the needs of scientific applications such as agricultural planning, POESIA combines ontologies, workflows and activity models to provide novel facilities for data integration using cooperative services. This approach pursues the vision of the Semantic Web [5, 6] and offers some concrete solutions for data integration and service composition on the Web. [3] describes the POESIA approach for organizing and reusing data and processes. It presents well-defined specialization and aggregation operators over activity patterns, allowing the adaptation of process frameworks for different situations. This approach also includes mechanisms to check the semantic consistency of compositions of Web services implementing activities and data repositories. [4] provides a more formal treatment of some aspects of the POESIA approach, such as structural and semantic properties of POESIA ontologies and implementation issues.

This paper introduces the POESIA ontological method for estimating data provenance. Domain ontologies depict the semantic relationships among terms, grouped according to different dimensions of one reality (e.g., geographic space or time). Tuples of terms, called *ontological coverages*, express the *scopes* of data sets and *granularities* of data values in these dimensions (e.g., the spatial extents and periods of time that a data set or value refers to). The semantic relationships between these terms induce a partial order among ontological coverages. This order is used to correlate scopes and granularities of data, enabling an estimation of data provenance. The major contribution of this paper is a framework for tracking data provenance, using ontologies to express data contents and the effect of chains of data integration operations on data sets. This framework can achieve efficient and fine grain provenance tracking.

The remainder of this paper is organized as follows. Section 2 presents an agricultural application used as a running example throughout the paper. Section 3 outlines the fundamentals of POESIA ontologies needed for provenance tracking. Section 4 describes the ontological method for estimating data provenance of aggregated values. Section 5 shows, from a more general perspective, how the ontological method for provenance estimation fits in the POESIA approach. It analyzes the role of domain ontologies with respect to both data integration and provenance tracking, in the context of inter-institutional workflows. Section 6 discusses related work. Finally, Section 7 summarizes contributions and extensions.

2. Motivating Example

The problem investigated here is the following: given a data item, what were the original data sources and the chain of data processing steps that produced it? Let us examine a real life scenario concerning data integration in agricultural applications. Figure 1(a) illustrates the consolidation of weather data through a hierarchy of intra and inter-institutional repositories, for use in agricultural planning. Each institution has a set of weather stations (data collecting devices), scattered across its operational area, to collect measurements such as maximum, minimum and average temperature and total rainfall per hour. These data are maintained in the repositories of the institutions that collect them. The spatial and temporal *scopes* of the institutional data sets (i.e., the land parcels and periods of time they cover) can overlap. For example, institution



(a) (b)

Figure 1: Integrating data sets in many steps

I1 operates in a limited region, while institution I2 has a wider spatial scope. Institution I3 encompasses units I3a and I3b. The data warehouse of consortium C1 consolidates data from I1 and I2, C2 from I2 and I3, and C3 from C1 and C2. This processing scheme produces data sets with increasingly broader scopes and denser sampling. The data granularity in the upper levels can be either the same or coarser than the granularity of the source data (e.g., from an hourly to a daily basis). The data at the lower levels tend to be more detailed and precise (but not necessarily accurate), while the data at the higher levels usually convey more abstraction, since they refer to increasingly broader scopes. Typical operations to produce such aggregations of the source data can be seen as variations of the basic data cube operations such as slice, dice, roll-up, and drill down.

Figure 1(b) gives a general view of the step-wise integration of weather data. First, the raw data collected by the weather stations of each institution are gathered, reviewed and stored as temporal series. Then, aggregation of historical data from each weather station generates the climate attributes for that particular point on the earth surface (e.g., average temperature and rainfall per month). Data warehouses (such as those in C1, C2 and C3) offer unified access to climate attributes originated from several sources, with aggregation and interpolation facilities for recovering consolidated data – typically OLAP to select and aggregate data over time and space, and interpolations to produce maps with estimations of the distribution of climate measurements across the lands. Finally, applications such as agricultural zoning [3] integrate and fuse data taken from these warehouses, among other sources, to derive other relevant information. Most of these applications need to understand not only the semantics of the data used, but also their provenance.

Figure 2 shows the star schema of the data warehouses used in the case studies throughout this paper. The Climate data warehouse has a data table with the values of maximum, minimum and average temperature and total rainfall, organized according to the dimensions of territorial divisions, time, agricultural products and organizations. The Crops production warehouse maintains data about the planted area, production, unit and monetary value, for each county, month and crop produced (minimum granularity this warehouse provides). These data warehouses can answer queries involving selection on the data they maintain and aggregation of the selected data to produce values of coarser granularities. For example, one can ask the Crop production warehouse for the total production of fruits, during the year 2001, in the whole Brazil, aggre-

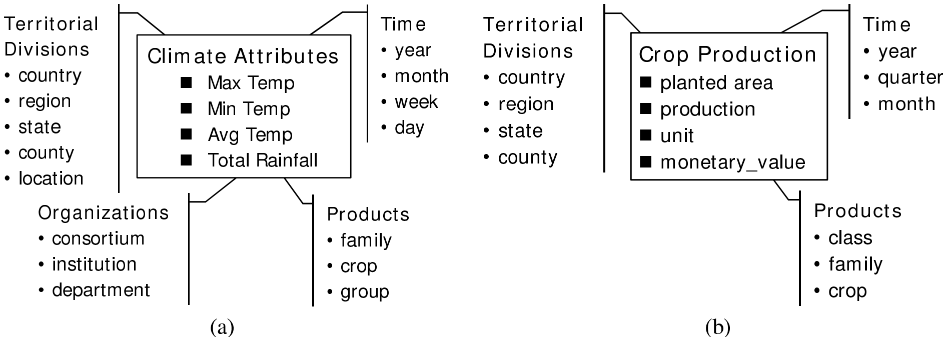
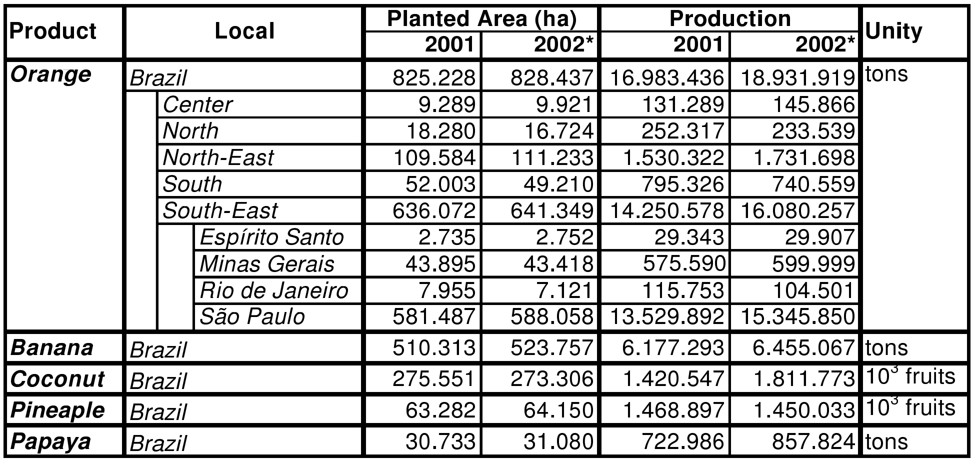


Figure 2: Agricultural data warehouses: (a) climate attributes; (b) crops production

gating the results for each fruit and state. Figure 3 presents an OLAP view of data provided by the Crops production warehouse[1]. This view shows the planted area and production of some fruits, detailing the granularity of the data about oranges for each Brazilian region and for each state of the region South-East.

Notice the similarities between the respective dimensions (territorial divisions, time and agricultural products), of the two data warehouses of Figure 2. The following sections show how to represent these dimensions in an ontology and the use of such an ontology to help track data provenance.



Product	Local			Planted Area (ha)		Production		Unity
				2001	2002*	2001	2002*	
Orange	*Brazil*			825.228	828.437	16.983.436	18.931.919	tons
		Center		9.289	9.921	131.289	145.866	
		North		18.280	16.724	252.317	233.539	
		North-East		109.584	111.233	1.530.322	1.731.698	
		South		52.003	49.210	795.326	740.559	
		South-East		636.072	641.349	14.250.578	16.080.257	
			Espírito Santo	2.735	2.752	29.343	29.907	
			Minas Gerais	43.895	43.418	575.590	599.999	
			Rio de Janeiro	7.955	7.121	115.753	104.501	
			São Paulo	581.487	588.058	13.529.892	15.345.850	
Banana	*Brazil*			510.313	523.757	6.177.293	6.455.067	tons
Coconut	*Brazil*			275.551	273.306	1.420.547	1.811.773	10^3 fruits
Pineaple	*Brazil*			63.282	64.150	1.468.897	1.450.033	10^3 fruits
Papaya	*Brazil*			30.733	31.080	722.986	857.824	tons

Figure 3: An OLAP view of the crops production data

3. POESIA Ontologies and Ontological Coverages

Figure 4 shows the organization of the space dimension as described in a POESIA ontology. The directed acyclic graph on the left, called an *arrangement of concepts*, formalizes the semantic relationships among the territorial subdivision concepts. The edges representing PART_OF relationships have a black circle close to the specific concept, and the edges representing IS_A

[1] This data has been gently provided by the personnel of the project Fruits Production Database from Embrapa Information Technology for Agriculture. The data for 2002 is an estimation, because the production assessment was not complete by the time this view was produced.

relationships have a diamond close to the component concept. This graph denotes that a Country is composed of a set of States or, alternatively, a set of Country Regions. A Country Region may be a Macro Region, an Official Region or another kind of region. Macro and Official Regions are composed of States, but a region of type Metro Area is composed of Counties. Eco Region and Macro Basin define other partitions of space, based on ecological and hydrological issues, respectively. Arrangements of concepts provide a general framework, being instantiated by arrangements of terms. The middle part of Figure 4 illustrates a subgraph of the arrangement of territorial subdivision concepts. An *arrangement of terms* instantiated from these concepts is represented by the directed acyclic graph (in this case a hierarchy) on the right side. There are also SYNONYM relationships not represented in the figure due to space limitations (e.g., BR can be used as a synonym to Brazil). An instantiated term needs to be qualified with the corresponding concept, in order to avoid ambiguity. Thus, State(RJ) refers to the state of Rio de Janeiro, while County(RJ) refers to the county of the same name.

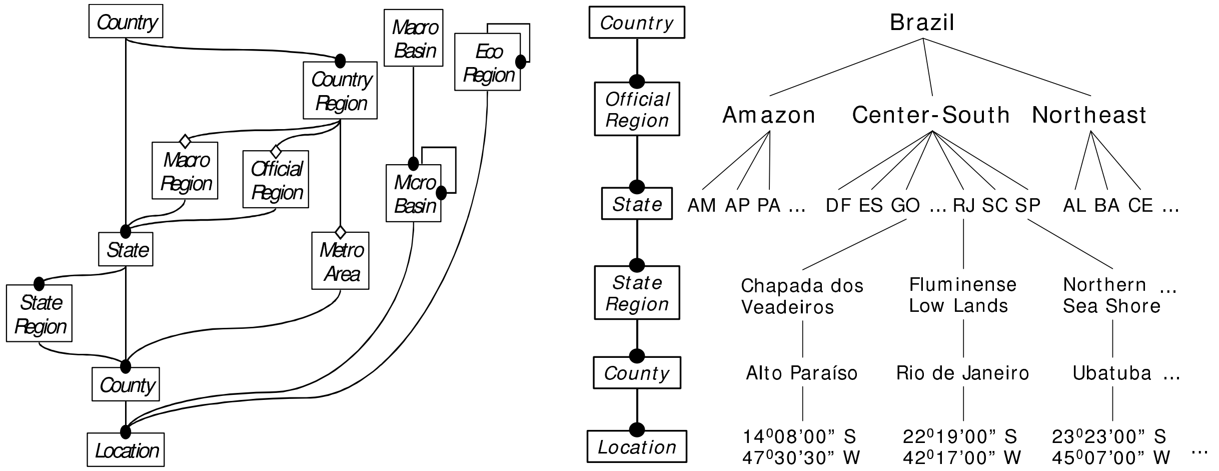


Figure 4: The space dimension of a POESIA ontology: (left) arrangement of concepts; (middle and right) a compatible arrangement of terms

Similar structures describe concepts and instantiated terms for other dimensions (such as time and products). The arrangements of concepts and terms for all the relevant dimensions constitutes a *POESIA ontology*. A tuple of terms from a POESIA ontology, called an *ontological coverage*, can describe the *scope* of a data set or the *granularity* of an aggregated value. For example, the ontological coverage [State(RJ),Crop(orange),Year(2002)] restricts the scope to the intersection of the spatial, crop and temporal scopes defined by the terms State(RJ), Crop(orange) and Year(2002) in a multidimensional space. The ontological coverage [State(RJ),State(SP)], on the other hand, denotes the union of the spatial scopes expressed by the two terms, because both refer to the same dimension. To narrow the scope in a particular dimension one has to choose a more specific term (e.g., go from State(SP) to County(Ubatuba)).

The semantic relationships represented in POESIA ontologies induce a partial order among ontological coverages that we call *semantic encompassing*. For example, the country called Brazil encompasses the state whose acronym is RJ (Rio de Janeiro), what is denoted by the expression [Country(Brazil)] $\models$ [State(RJ)]. Furthermore, [State(RJ)] $\models$ [State(RJ),Year(2002)] and [State(RJ),State(SP)] $\models$ [State(SP)].

A data set can be associated with an ontological coverage expressing its application scope, i.e., the scope it has data about (e.g., a particular state such as `State(RJ)`). A data value can be associated with an ontological coverage expressing its granularity (e.g., `[State(Brazil), Year (2002)]` to refer that a data value refers to the whole Brazil during 2002). The scope of a data set must encompass the scopes of its components and the minimal granularities of the data values contained in this data set. The scope of a data value is equivalent to its granularity. Two ontological coverages are *equivalent* if they refer to the same scope (e.g., `[Country(BR)]` $\equiv$ `[Country(Brazil)]`. We can show, for a limited set of semantic relationships between terms, that the encompassing relationship is reflexive and transitive. The equivalence relationship is also symmetric. A more complete and formal treatment of POESIA ontologies, with demonstration of properties, definition of consistency rules and description of their multiple applications can be found in [4].

4. Ontological Estimation of Data Provenance

Let us consider the union of data sets in data warehouses. The ontological coverages described in the previous section can express the scope of the data sources and of the resulting data sets. Figure 5(a) illustrates the data flow for the consolidation of crop production data, involving cooperating institutions and consortia. The scopes of the data repositories are described by the ontological coverages attached to the nodes. For instance, institution `I1` maintains data about the production of grains in the center-south region of Brazil during the year 2002, while `I2` is concerned with the production of fruits in the whole of Brazil during the same year. The information flow, indicated by the arrows, shows for example that the data set of consortium `C1` consolidates data from `I1` and `I2`, in a scope encompassing those of its sources: the production of food in Brazil during 2002.

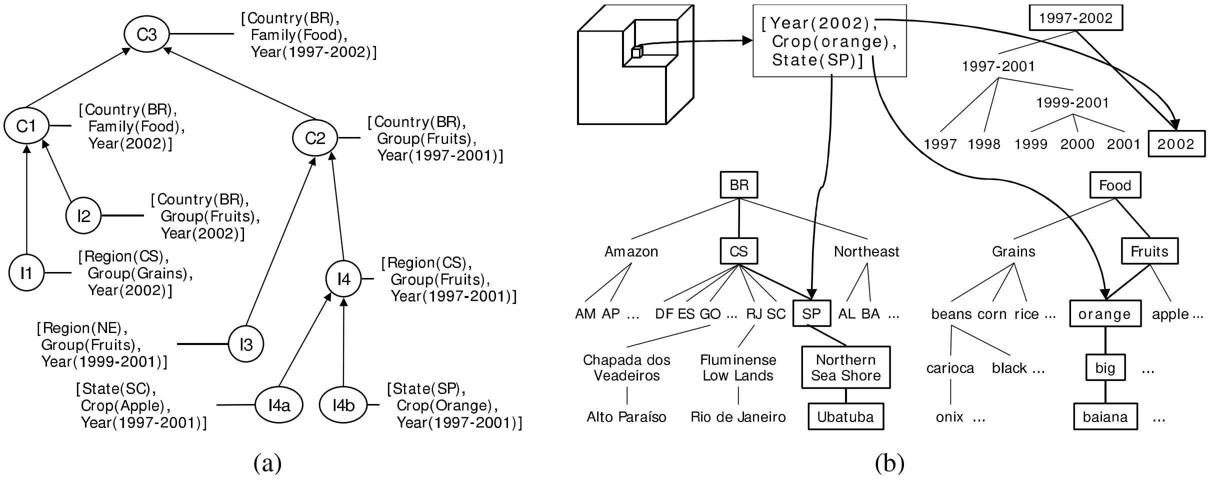
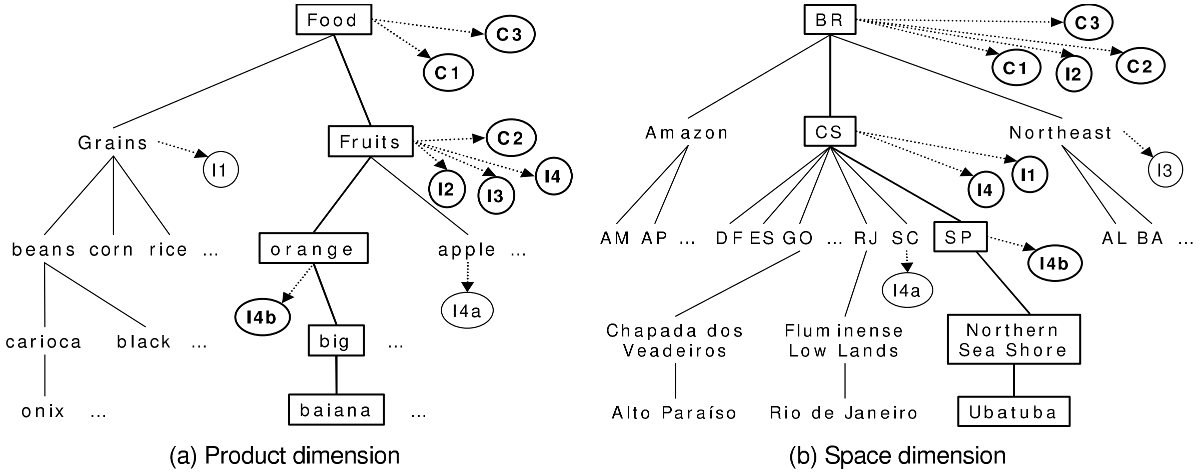

C3
[Country(BR), Family(Food), Year(1997-2002)]
C1
[Country(BR), Family(Food), Year(2002)]
C2
[Country(BR), Group(Fruits), Year(1997-2001)]
I2
[Country(BR), Group(Fruits), Year(2002)]
I1
[Region(CS), Group(Grains), Year(2002)]
I4
[Region(CS), Group(Fruits), Year(1997-2001)]
[Region(NE), Group(Fruits), Year(1999-2001)]
I3
[State(SC), Crop(Apple), Year(1997-2001)]
I4a
I4b
[State(SP), Crop(Orange), Year(1997-2001)]
(a)
[Year(2002), Crop(orange), State(SP)]
1997-2002
1997-2001
1999-2001
1997 1998 1999 2000 2001 2002
BR
Food
Amazon
CS
Northeast
Grains
Fruits
AM AP ...
DF ES GO ... RJ SC
SP
AL BA ...
beans corn rice ...
orange
apple ...
Chapada dos Veadeiros
Fluminense Low Lands
Northern Sea Shore
carioca black ...
big
...
Alto Paraíso
Rio de Janeiro
Ubatuba
onix ...
baiana
...
(b)

Figure 5: The use of POESIA ontologies: (a) scopes of cooperating services; (b) the level of granularity of an aggregated value

The provenance of an aggregated value in a node can be estimated by analyzing the scopes of the data sources of the node. The potential sources, for each dimension, are those whose ontological coverage overlaps (encompasses or is encompassed by) the coverage of the aggregated value in that dimension. For example, consider the average production of oranges in São Paulo

State during 2002. Figure 5(b) shows how the ontological coverage expresses the granularity of the aggregated value, by indicating specific terms in different dimensions of a POESIA ontology. Each term whose semantics overlaps the ontological coverage of the aggregated value is surrounded by a rectangle.

Figure 6 illustrates the identification of the potential data sources for different dimensions. It shows the arrangements of concepts for the space and product dimensions, with pointers associating the data sources with the terms used to express their scopes (e.g., C3 is associated with BR because its ontological coverage refers to Country(BR)). Then, provenance tracking in one dimension reduces to collecting the sources associated with all the ancestors and descendants of the terms expressing the coverage of the aggregated value in that dimension. Figure 6(a) highlights the potential sources in the space dimension. For instance, sources C3, C1, C2 and I2 are candidates because their ontological coverages refer to Country(BR) and Country(BR) $\models$ State(SP). I4b is also a potential source because its ontological coverage refers directly to State(SP). If there have been other sources associated with the descendants of State(SP) they should also be taken into account. I3 and I4a are not potential sources because they refer to nodes outside of the closure of ancestors and descendants of State(SP). Figure 6(b) shows the same method applied to the product dimension. A similar analysis can be done for the time dimension.

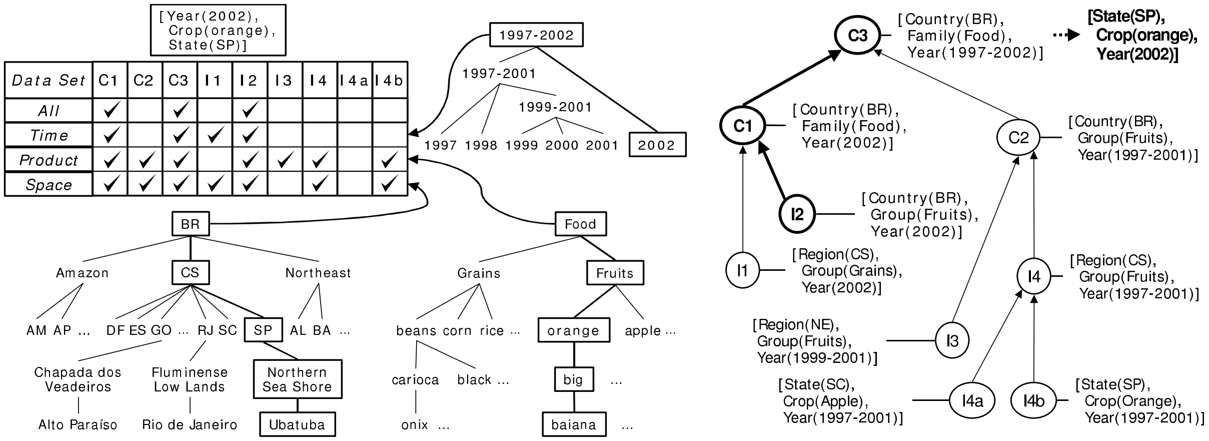


(a) Product dimension

(b) Space dimension

Figure 6: Potential data sources for the production of oranges in São Paulo State

The potential sources for an aggregated value are those figuring as candidates in all dimensions. Figure 7 illustrates the ontological scheme for the estimation of the data provenance. The table on the left side shows that only C1, C3 and I2 figure as potential sources in all dimensions. We call the subgraph (or sub-workflow) obtained by considering only the potential sources and the links between them the *relevant flow* for the data provenance of the given data item. It shows all the possible flows of data from the potential sources to the data item. Figure 7(b) highlights the relevant flow for the aggregated value considered. The granularity of that value, expressed by [State(SP),Crop(orange),Year(2002)], can be used to select the specific data items which may have been used to calculate the aggregation. This method gives only an estimation of the data provenance because the overlapping of the scopes of the data sources can lead to alternative paths for supplying a particular data value.



Data Set	C1	C2	C3	I1	I2	I3	I4	I4a	I4b
All	✓		✓		✓				
Time	✓		✓	✓	✓				
Product	✓	✓	✓		✓	✓	✓		✓
Space	✓	✓	✓	✓	✓		✓		✓

Figure 7: Appraising data provenance: (a) contrasting dimensions; (b) estimated data flow

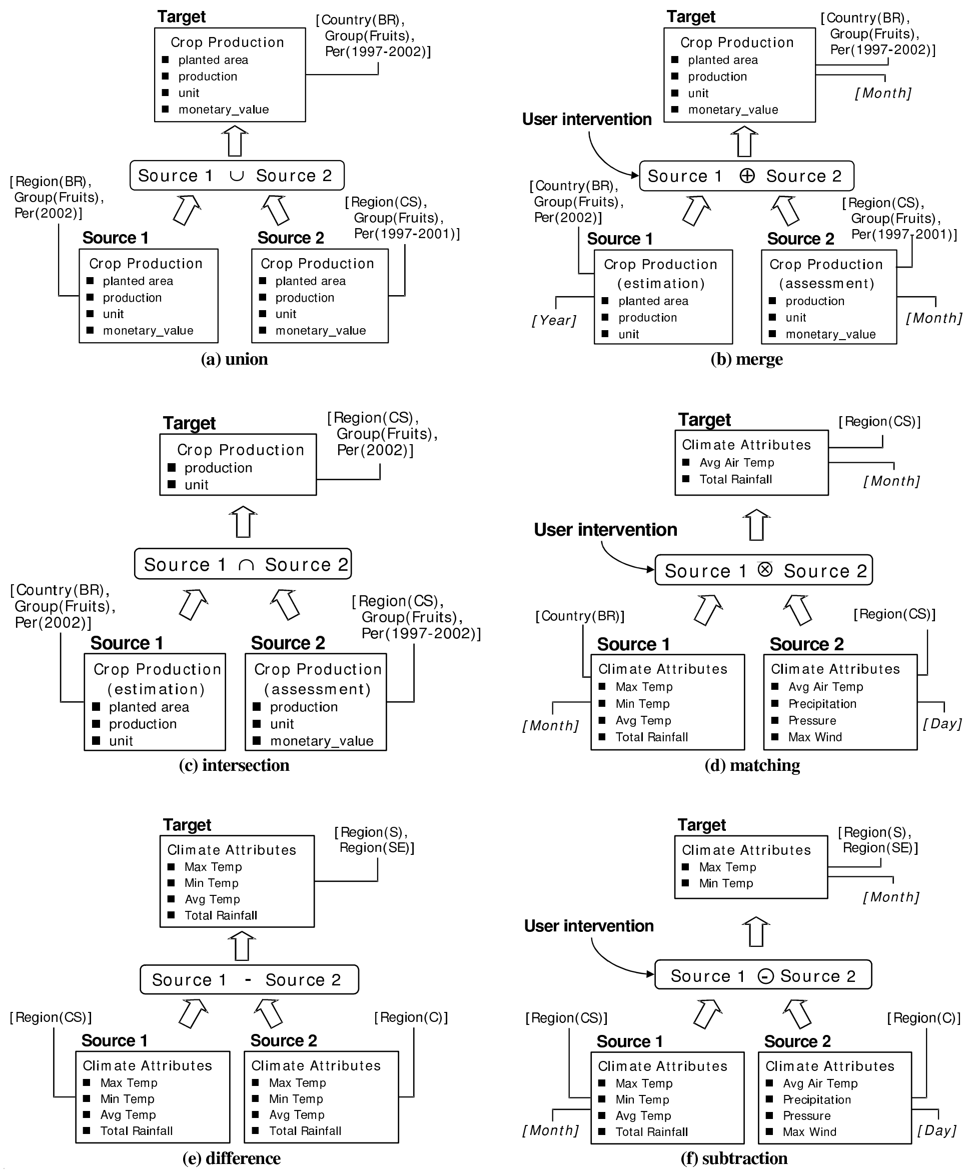
5. Ontological Nets for Data Integration

An *ontological net for data integration* is an infra-structure for consolidating and fusing data through distributed cooperative processes, where the description, discovery and composition of data sets and services are based on domain ontologies. In order to better explain this concept and how the ontological method for provenance tracking fits in the POESIA approach, let us analyze the basic operators for data integration in cooperative geographical applications and the role of domain ontologies in this context, from a higher level perspective.

5.1. Data Integration Operators

The POESIA approach classifies the operators typically used for integrating data in cooperative geographical applications in three categories: combination of data sets, filtering data and transforming data values. Figure 8 presents some examples of the operators for combining data sets. The `union` operator collects data items from two data sources into a composite data set, whose schema matches those of the sources. In Figure 8(a), data about the production of fruits in Brazil during 2002 is united with another data set about the production of fruits in the Center-South region of the country between 1997 and 2001, generating a data set which covers the production of fruits in Brazil from 1997 to 2002. The `merge` operator relaxes the semantics of the union operator by allowing slightly different semi-structured data sources and user intervention to solve conflicts. Figure 8(b) shows an example of merging two heterogeneous data sets, into a semi-structured data set, whose schema is a composition of the source schemas. The `union` and `merge` operators produce data sets whose scope encompasses those of the data sources. The result may contain data with the granularities present in both sources. Additionally, POESIA ontologies help to identify conflicts on merging data sets in the absence of a common key. Data items from different sources, but with equivalent scopes, called *semantically identifiable matches*, are converted into one item in the target, using heuristics and user intervention to solve conflicts. For example, one can detect discrepant values between data items referring to the same product, at the same place and time and put only the correct value in the target.

The `intersection` operator employs heuristics to produce data items in the target for



(a) union **(b) merge**

(c) intersection **(d) matching**

(e) difference **(f) subtraction**

Figure 8: Combining data sets

each pair of matching items from the two data sources. The schema of the target can be the union or the intersection of the source schemas, depending on the matching data items. Figure 8(c) shows the intersection of two heterogeneous data sets about crop production. The matching operator is similar to the intersection, but allows user intervention to analyze matches and define the target schema. For example, one can identify that Total rainfall in Source 1 matches Precipitation in Source 2, define the corresponding target attribute and choose the data values to put in the target. Figure 8(d) shows the matching of two heterogeneous sources of weather data. For intersection and matching, the scope of the target data set is the intersection of those of the data sources, and the minimum granularity provided by the target is the maximum among the minimum granularities of the sources.

The difference and the subtraction operators return the data items of the first data source which do not have a match in the second source. The resulting schema derives from the schema of the first data source. The difference between these operators is that subtraction allows heterogeneous schemas and user intervention. Figures 8(e) and 8(f) illustrate the application of these operators to climate data sets. For both operators the scope and the minimum granularity of the target is given by subtracting the scope and minimum granularity of the second source from those of the first one.

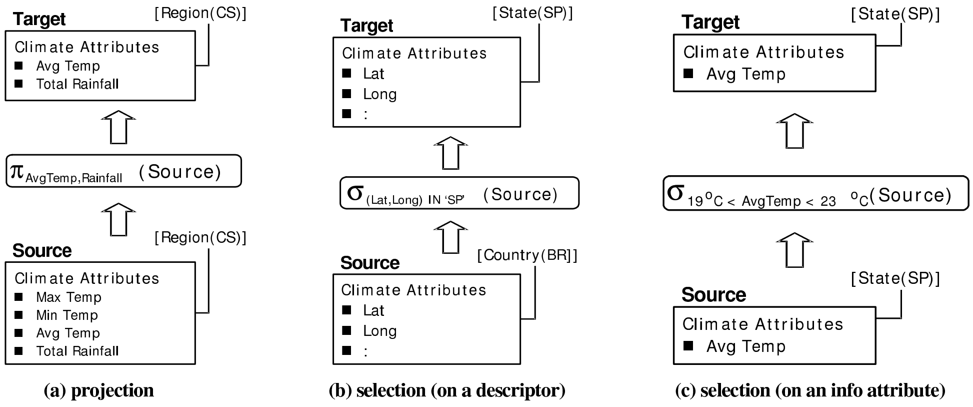


Figure 9: Data Filtering

Figure 9 illustrates the operators for filtering data sets: projection and selection. These operators keep the semantics of the corresponding relational operators, i.e., projecting attributes or selecting data items according to some predicate, respectively. Projection preserves the scope of the source in the target (Figure 9(a)), while selection may not. If the selecting predicate stipulates filtering on a term defined in a POESIA ontology, the restricted scope of the target can be determined by that term (Figure 9(b)). However, this is not so straightforward for filtering on values of the data table (Figure 9(c)).

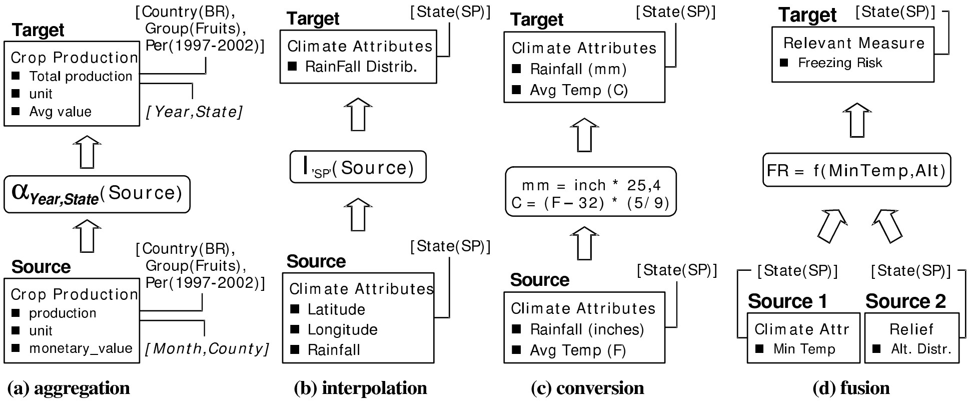


Figure 10: Transforming values

Figure 10 presents the operators that transform data values. The aggregation calculates coarse grain measurements from data in finer granularities. Figure 10(a) illustrates the aggregation of crop production data for each month and county into the respective values for each year and state. The interpolation estimates the continuous distribution of measurements from discrete samples. Figure 10(b) illustrates the interpolation of average rainfall samples to produce a map expressing the distribution of this measurement across the lands. The `conversion` employs user defined functions to convert data (e.g., from one measurement unit into another). Figure 10(c) illustrates the conversion of rainfall measurements from inches to millimeters and measurements of average temperature, for the same scope, from Fahrenheit to Celsius degrees. Finally, the `fusion` operator combines values from different data sources, whose respective scopes match each other, into another meaningful measurement, according to user defined functions. Figure 10(d) illustrates the synthesis of the freezing risk from the minimum temperature and altitude. All these operators preserve the scopes of the data sets, though only `aggregation` and `interpolation` impact the data granularity.

5.2. Data Reconciling through Articulation of Ontologies

POESIA ontologies help the integration process with respect to data scopes and granularities as discussed in Section 5.1. General and application ontologies help to investigate the semantic correspondences among heterogeneous data items and index libraries of data conversion functions. Some decisions made when integrating data must be annotated, in order to explain the relevant details of data provenance that cannot be captured by ontological coverages alone.

Let us consider the integration of two heterogeneous data sets of weather measurements from distinct institutions, in a particular portion of a cooperative process. The schema for the semi-structured data of each data set can be represented as a directed graph (e.g., using XML). The POESIA approach enriches these graphs with metadata describing the data elements, and uses ontologies to express the properties of these elements and interrelate them. These enriched schemas are themselves specific ontologies. Thus, ontologies articulation [7] can be used as a basis to integrate data sources. Figure 11 illustrates this approach. The two graphs at the bottom of the figure describe the data sources, the graph at the top represents the target data set and the dotted and dashed links between nodes of these graphs represent the articulation rules, i.e., the data flows from the sources to the target. These articulations show, for example, that the values of latitude and longitude from the source in the left-bottom corner of the figure, represented in degrees, minutes and seconds, must be converted into degrees and decimals of degrees to be inserted into the target.

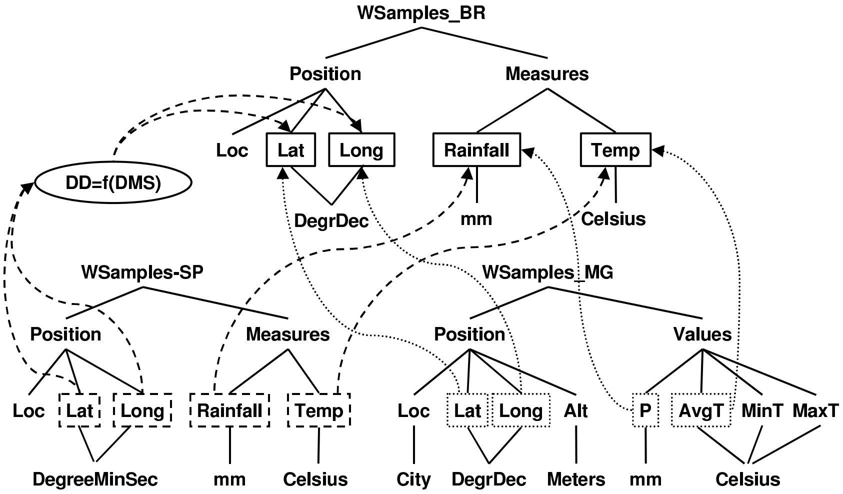
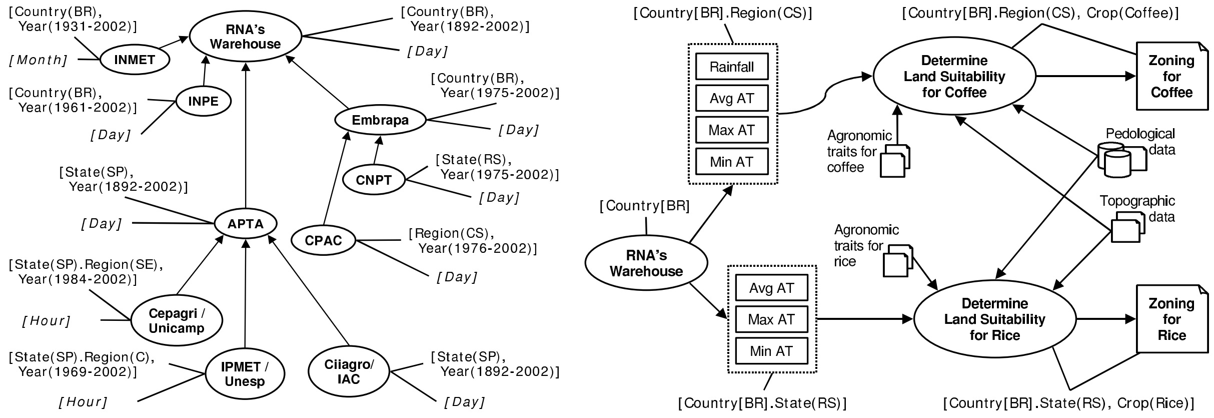


Figure 11: Reconciling heterogeneous data sets by ontologies articulation

In summary, many kinds of ontologies are necessary for reconciling heterogeneous data sets. Application ontologies, derived from heterogeneous schemas, describe the data sets to be integrated. General ontologies, such as Wordnet and ontologies of measurement units, help to establish the correspondences and conversion procedures among data items. Finally, domain ontologies and ontological coverages help to identify correspondences according to scopes or granularities of heterogeneous data in several dimensions, such as space, time and categories of products. The correspondences among data items of different data sets can be seen as articulations between application ontologies, as illustrated in figure 11.

5.3. Semantic Workflows

Semantic workflows are cooperative process running on ontological nets. These processes employ data integration operators, according to ontologies articulations. POESIA ontologies contribute to render a general view of what is going on in these workflows, by expressing the scopes and granularities of the data involved. Figure 12(a) illustrates the integration of weather data from different institutions. Each service is characterized by its scope and the minimum granularity it supports for data recovery. For example, the `INMET` (Instituto Nacional de Meteorologia – National Institute of Meteorology) collected weather data samples across Brazil in the period between 1931 and 2002. The minimum time granularity for the data supplied by `INMET` is month. The ultimate recipient of data in this cooperative process is the `RNA Warehouse` (Rede Nacional de Agrometeorologia – National Agrometeorology Network), which can provide weather and climate data about virtually any place in Brazil. The temporal scope of the weather data supplied by the `RNA Warehouse` is 1892 to 2002 and the minimum granularity supported is day. The granularity for each data item depends on the sources of that item. Figure 12(b) illustrates the role of the `RNA Warehouse` on supplying climate data to determine land suitability for different crops. The scope of the sub-processes for determining land suitability for coffee and rice must be compatible with the coverages of the respective sub-sets of climate attributes recovered from the `RNA Warehouse` (see [3] for details).



[Country(BR), Year(1931-2002)]
[Month]
INMET
RNA's Warehouse
[Country(BR), Year(1892-2002)]
[Day]
[Country(BR), Year(1961-2002)]
INPE
[Day]
[Country(BR), Year(1975-2002)]
Embrapa
[Day]
[State(RS), Year(1975-2002)]
CNPT
[Day]
[State(SP), Year(1892-2002)]
[Day]
APTA
CPAC
[Region(CS), Year(1976-2002)]
[Day]
[State(SP).Region(SE), Year(1984-2002)]
[Hour]
Cepagri / Unicamp
[State(SP).Region(C), Year(1969-2002)]
[Hour]
IPMET / Unesp
Ciiagro/ IAC
[State(SP), Year(1892-2002)]
[Day]
[Country[BR].Region(CS)]
Rainfall
Avg AT
Max AT
Min AT
[Country[BR].Region(CS), Crop(Coffee)]
Determine Land Suitability for Coffee
Zoning for Coffee
Agronomic traits for coffee
Pedological data
Topographic data
[Country[BR]
RNA's Warehouse
Agronomic traits for rice
Avg AT
Max AT
Min AT
Determine Land Suitability for Rice
Zoning for Rice
[Country[BR].State(RS)]
[Country[BR].State(RS), Crop(Rice)]

(a)

(b)

Figure 12: Ontologies as a framework for estimating data provenance: (a) scopes and minimum granularities of cooperating services; (b) the use of the integrated data by different processes

6. Related Work

The traditional solutions for tracking data provenance, some of which consider general data formats and processing, employ metadata to annotate the processing history [8, 9, 10, 11]. However, these solutions do not scale well to large data sets, long processing flows and fine grained provenance. Many other studies on data provenance are limited to views defined by query operations on databases, calling this restricted problem *lineage tracing*. Woodruff *et al.* [12] introduce the concept of *inverse query*, which maps an output to the data items used to produce that output. They define the class of functions admitting inversion and the concept of *weak inversion* to estimate the lineage for a wider class of functions. However, they do not show how to determine the inverse queries, but expect the data transformation definer to provide them.

Cui *et al.* [13, 14] define the *lineage* of the result of a relational database query as the minimal set of tuples necessary to produce that result. They present an algorithm for tracing lineage over chains of `aggregate-select-project-join` views. Their approach is based on the inversion of the view definition and requires materializations of original relations and intermediate views. [15] generalizes their previous results for graphs of general transformations used for loading data warehouses. Nevertheless, their methods are built upon some constraints and specific information about the sources and transformations employed, and require considerable storage for intermediate results.

Buneman [1] distinguishes between *why provenance* and *where provenance*. The former refers to the data items which have some influence on the result (e.g., which determine the logical value of a predicate used to select tuples). The latter refers to the items effectively used to synthesize the result (e.g., multiple values summed up to obtain an aggregated value like average). He provides a framework to track both kinds of data provenance for specific classes of `select-project-join-union` queries in a data model generalizing relational and hierarchical data representations such as XML. Galhardas *et al.* [16] present some data lineage facilities coupled to a data cleansing scheme based in a graph of transformations with exceptions management to support the refinement of the cleaning criteria. Fan [17] provides algorithms to trace data lineage in automatically reversible sequences of schema conversions, employing the hyper-graph based high level data model and the functional query language of the Automed system.

Therefore, current approaches either support just coarse grain provenance tracking or rely on detailed descriptions of the data sources and the data transformations applied (e.g., schemas and query expressions), making them unfit in many situations for cooperative systems over the Web. Furthermore, these approaches lack abstraction mechanisms to enable a general understanding and exploration of the information flow. To the best of our knowledge, the notion of domain ontologies [18, 19, 20] has not been yet exploited as a framework for tracking data provenance. This paper has shown that such a solution can eliminate some of these shortcomings.

7. Conclusions

Data provenance tracking is becoming increasingly important as more on-line data sources become available. This paper has shown how domain ontologies are used in POESIA as a basis for tracking data provenance in cooperative processes involving data integration. POESIA employs tuples of domain specific terms defined in multidimensional ontologies to correlate the scope and granularities of the target data with those of the data sources, enabling the estimation of the data provenance. Additionally, POESIA ontologies help to semantically identify matches on heterogeneous data sources, i.e., data items from different sources referring to the same scope. It helps to detect and solve conflicts among heterogeneous data sources, and allows tracking the data transformation flow across chains of data integration operators.

The benefits of this ontological method for estimating data provenance are (1) a framework for understanding data provenance based on domain specific concepts; (2) support for fine grain provenance tracking; (3) precision and conciseness for expressing the scopes and granularities; (4) coupling with an approach for data integration and services composition; (5) the cost for maintaining the infra-structure for provenance tracking is shared with facilities for cataloging, discovering and integrating data and services.

This research has been focused on the conceptual definition and formalization of the POESIA approach. Ongoing work includes the implementation of prototypes to evaluate this approach on scientific applications in agriculture. We have developed an ontology for this application domain, and a Java package for importing and using domain ontologies in application programs. This package includes classes for coping with ontological coverages: selecting them from the ontology and comparing them for encompassing and equivalence relationships. Now, we are incorporating this package into the WOODSS system for decision support based on scientific workflows [21].

The evaluation of the ontological method for tracking data provenance depends on the availability of case studies developed in accordance with the POESIA approach, i.e., a collection of interconnected data sets with the associated ontological coverages. In addition, comparing the ontological method with other approaches for provenance tracking will require the development of new criteria for taking measures and comparing results, as the ontological method leads to a different balance of costs and benefits, such as the cost for developing domain ontologies versus the benefits of their use for data integration, data and processes reuse and provenance tracking. Other extensions of this work are: developing optimal algorithms for provenance tracking by the ontological method, making experiments to asses space and time complexity and the effectiveness of the ontological method in practice, and compare the results with those of the traditional approaches.

Acknowledgments

The authors from Campinas University are partially supported by Embrapa, CAPES, CNPq and the Finep/Pronex/IC/SAI95/97 project. The authors from Georgia Tech are partially supported by two grants from the Operating Systems and ITR programs (CISE/CCR division) of NSF, by a contract from the SciDAC program of DoE, and a contract from the PCES program (IXO) of DARPA. The application scenarios and data used in this work were ultimately provided by Brazilian experts in agriculture. Special thanks to Daniel Andrade and Flavio Silva from the IC-Unicamp DB-group for the valuable discussions on preliminary versions of this material.

References

[1] P. Buneman, S. Khanna, and W. C. Tan. Why and where: A characterization of data provenance. In *Proc. Intl. Conf. on Data Theory (ICDT)*, volume 1973 of *LNCS*, pages 316–330. Springer, 2001.

[2] R. Bose. A conceptual framework for composing and managing scientific data lineage. In *Proc. Conf. on Statistical and Scientific Database Management*, pages 15–19. IEEE, 2002.

[3] R. Fileto, L. Liu, C. Pu, E. D. Assad, and C. B. Medeiros. POESIA: An ontological workflow approach for composing web services in agriculture. *The VLDB Journal*, 2003. (accepted for publication).

[4] R. Fileto. *POESIA: An Ontological Approach for Data and Services Integration on the Web*. PhD thesis, Institute of Computing, University of Campinas, Brazil, 2003.

[5] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, May 2001.

[6] W3C's Semantic Web Activity. `http://www.w3.org/2001/sw/` (as of July 2003).

[7] P. Mitra, G. Wiederhold, and M. L. Kersten. A graph-oriented model for articulation of ontology interdependencies. In *Proc. EDBT Conf.*, volume 1777 of *LNCS*, pages 86–100. Springer, 2000.

[8] D. P. Lanter. Design of a lineage-based metadata base for GIS. *Cartography and Geography Information Systems*, 18(4):255–261, 1991.

[9] P. Brown and M. Stonebraker. BigSur: A system for the management of earth science data. In *Proc. VLDB Conf.*, pages 720–728. Morgan Kaufmann, 1995.

[10] T. Lee, S. Bressan, and S. E. Madnick. Source attribution for querying against semi-structured documents. In *Proc. Workshop on Web Information and Data Management*, pages 33–39, 1998.

[11] P. A. Bernstein and T. Bergstraesser. Meta-data support for data transformations using microsoft repository. *IEEE Data Engineering Bulletin*, 22(1):9–14, 1999.

[12] A. G. Woodruff and M. Stonebraker. Supporting fine-grained data lineage in a database visualization environment. In *Proc. Intl. Conf. on Data Engineering (ICDE)*, pages 91–102. IEEE Computer Society, 1997.

[13] Y. Cui and J. Widom. Practical lineage tracing in data warehouses. In *Proc. Intl. Conf. on Data Engineering (ICDE)*, pages 367–378. IEEE, 2000.

[14] Y. Cui, J. Widom, and J. L. Wiener. Tracing the lineage of view data in a warehousing environment. *ACM TODS*, 25(2):179–227, 2000.

[15] Y. Cui and J. Widom. Lineage tracing for general data warehouse transformations. In *Proc. VLDB Conf.*, pages 471–480. Morgan Kaufmann, 2001.

[16] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C. Saita. Improving data cleaning quality using a data lineage facility. In *Proc. Conf. on Data Management and Data Warehouses (DMDW)*, volume 39 of *CEUR-WS.org*, 2001.

[17] H. Fan and A. Poulovassilis. Tracing data lineage using schema transformation pathways. In *Workshop on Knowledge Transformation for the Semantic Web KTSW/ECAI*, volume 95, pages 64–79. IOS Press, 2003.

[18] M. Uschold and M. Gruninger. Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155, 1996.

[19] N. Guarino. Formal ontology and information systems. In *Proc. Intl. Conf. on Formal Ontologies in Information Systems (FOIS)*, pages 3–15. IOS Press, 1998.

[20] M. Missikoff, R. Navigli, and P. Velardi. The Usable Ontology: An Environment for Building and Assessing a Domain Ontology. In *Proc. Intl. Conf. on Semantic Web*, volume 2342 of *LNCS*, pages 39–53. Springer, 2002.

[21] L. A. Seffino, C. B. Medeiros, J. V. Rocha, and Bei Yi. WOODSS - a spatial decision support system based on workflows. *Decision Support Systems*, 27(1-2):105–123, 1999.