Towards Best-Effort Merge of Taxonomically Organized Data

David Thau [#1], Shawn Bowers [*2], Bertram Ludäscher [#3]

[#] *Department of Computer Science, University of California*
Davis, USA
[1] thau@ucdavis.edu
[3] ludaesch@ucdavis.edu

[*] *Department of Computer Science, Gonzaga University*
Spokane, USA
[2] bowers@gonzaga.edu

***Abstract*— We consider the task of merging datasets that have been organized using different, but aligned taxonomies. We assume such a merge is intended to create a single dataset that unambiguously describes the information in the source datasets using the alignment. We also assume that the merged result should reflect the observations of the datasets as specifically as possible. Typically, there will be no single merge result that is both unambiguous and maximally specific. In this case, a user may be provided with a set of possible merged datasets. If the user requires a single dataset, that dataset loses specificity. Here we examine whether the data exchange setting can provide a way to derive a "best-effort" merge. We find that the data exchange setting might be a good candidate for providing the merge, but further research is needed.**

I. Introduction

In *data exchange*, a source schema $\mathbf{S}$ and a target schema $\mathbf{T}$ are given, together with source-to-target dependencies Σ_{st}, and target schema constraints Σ_t. Given an input instance I of $\mathbf{S}$, a *solution* to the data exchange problem is a target instance J of $\mathbf{T}$, such that $\langle I, J \rangle$ satisfies Σ_{st} and J satisfies Σ_t. In general, there are multiple solutions, so the *certain* answers, i.e., contained in all possible solutions J, are usually reported. Data exchange has been well studied in recent years [1], [2], [3] and tractable algorithms for many common scenarios have been developed. In our application projects, we are interested in another problem, i.e., of *merging taxonomically organized datasets*, using articulations between taxonomies. In this paper we present some first ideas for casting the latter problem as a variant of the former, with the hope that we can leverage existing results in data exchange for our data merge problem.

A. Merging Taxonomically Organized Datasets

We address the problem of merging datasets that have been registered to different, but aligned taxonomies (inheritence hierarchies). Consider, e.g., the simplistic scenario in Fig. 1: nodes A, B, C are *concepts* from taxonomy T_1, while D is from T_2. As usual, concepts denote sets (whose members may be unknown) and thus can be represented by unary relations. Within a taxonomy, an arrow $\mathsf{A} \rightarrow \mathsf{C}$ denotes an inclusion (or *isa*) relation; its semantics is captured via a first-order statement $\forall x : \mathsf{A}(x) \rightarrow \mathsf{C}(x)$. When merging datasets that have been registered to *different* taxonomies T_1 and T_2, we assume that we have at least partial information in the form of *articulation constraints* between pairs of concepts from T_1 and T_2.[1] For example, the articulation $\mathsf{T}_1.\mathsf{B} \subsetneq \mathsf{T}_2.\mathsf{D}$ (or $\mathsf{B} \subsetneq \mathsf{D}$ for short) states that (i) every (member of) B is also a (member of) D, and (ii) some D are not B, i.e., $\mathsf{D} \setminus \mathsf{B}$ is not empty. Similarly, the articulation $\mathsf{A} \oplus \mathsf{D}$ states that concepts A and D *overlap*, i.e., the sets $\mathsf{A} \setminus \mathsf{D}$, $\mathsf{D} \setminus \mathsf{A}$, and $\mathsf{A} \cap \mathsf{D}$ are all non-empty. As articulation constraints, we use the mutually exclusive RCC-5 relations $\oplus$, $\subsetneq$, $\supsetneq$, $!$, $\equiv$, denoting overlap, proper part and its inverse, disjointness, and equivalence, respectively [4]. These relations have proven to be useful in reasoning about taxonomies and taxonomy alignments [5], and can be used to merge taxonomies into a single, combined taxonomy [6].

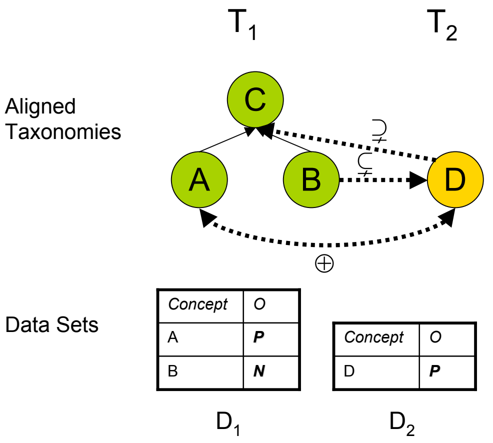
In addition to the taxonomies and articulations, Figure 1 describes two datasets D_1 and D_2. In this paper, we restrict ourselves to very simple datasets that describe the *presence* or *absence* of instances of named concepts at a single place and time. For example, the tuple $(\mathsf{A}, \mathtt{P}) \in D_1$ states that at least one instance of the concept (or taxon) $\mathsf{T}_1.\mathsf{A}$ was observed; $(\mathsf{B}, \mathtt{N}) \in D_1$ means that *no* instances of $\mathsf{T}_1.\mathsf{B}$ were observed.[2] Similarly, D_2 states that at least one instance of $\mathsf{T}_2.\mathsf{D}$ was observed. Note that a given dataset may be logically inconsistent with its associated taxonomy. For example, a dataset annotated to T_1 would be inconsistent if it reported the presence of A but the absence of C (since every A is also a C). Presence/absence datasets such as these are commonly used in ecology, biodiversity, and evolutionary biology research. For example, knowledge about where a given species was observed can be used to predict other locations where it may be found. This type of analysis, called species niche modeling, has been used to predict where invasive species are likely to take root [7], how diseases are likely to spread [8], and how global warming may impact biodiversity [9]. Many online databases of species occurrence data exist.[3] If all of the databases used the same

[1] Datasets using the same taxonomy can be merged by unioning them.

[2] $\mathtt{P}$ and $\mathtt{N}$ stand for *present* and *not present*, respectively.

[3] As of the date of this publication, the Global Biodiversity Information Facility Data Portal (http://www.gbif.org) provides access to almost 190 million occurrence observations, most of which have been georeferenced.

978-1-4244-6523-1/10/$26.00 © 2010 IEEE



Concept	O
A	***P***
B	***N***

Concept	O
D	***P***

Fig. 1: Aligned taxonomies T_1, T_2 with datasets to be merged.

taxonomy to organize and name species, the occurrence data from these databases could be easily combined. Unfortunately, there frequently exist many taxonomies organizing a given set of species, and the meaning of a given taxon name may differ depending on which taxonomy is used [10]. For example, two observed organisms may be given the same species name according to one taxonomy, but different species names according to a second taxonomy. To address the problems this causes for data integration, experts create alignments between frequently used taxonomies [11], [12], [13].

Given this setting, it is natural to ask how two datasets D_1 and D_2 can be combined into a single merged dataset D_3 annotated to a new taxonomy T_3 created by merging the initial pair of taxonomies $\mathsf{T}_1, \mathsf{T}_2$. We call this setting *taxonomic data merge*. The taxonomic data merge setting can be seen as a variant of the standard data exchange scenario as follows: The union of datasets $D_1 \cup D_2$ constitute the source instance I, the concepts in the taxonomies $\mathsf{T}_1, \mathsf{T}_2$ become relations in the source schema $\mathbf{S}$, the concepts in the merged taxonomy T_3 make up the relations in the target schema $\mathbf{T}$, the source-to-target constraints Σ_{st} are somehow derived from articulations/taxonomy constraints, and the relationships in the merged taxonomy are represented by Σ_t. The solutions J of this data exchange problem can then be understood as the solutions to our taxonomic data merge problem. There are, however, points of mismatch that complicate the comparison.

B. Dataset Constraints

Typically, when a dataset uses a concept drawn from a taxonomy, not every implied concept is included in the dataset. For example, D_1 in Figure 1 states the presence of concept A. We know from the semantics of presence and the taxonomy that the presence of C is implied by the presence of A. However, concept C does not appear explicitly in the dataset. Thus in our setting, one could argue that the datasets do not strictly adhere to some source schema constraints.

If, on the other hand, our datasets *did* follow this source schema constraint, we would have a different sort of problem. Imagine that someone using T_1 reported the presence of both A and C. Given no other information, this dataset has some ambiguity: Is the C that was reported a consequence of the A that was reported, or was there some other instance that was a C but not an A? Assuming that concepts A and C are not equivalent (which may be inferred from B), we can assume that there exist instances that are in C but not in A, i.e., $\mathsf{C} \setminus \mathsf{A}$ is not empty. The ambiguity in the dataset arises from the fact that we do not know whether or not such a thing was being reported in the dataset.

Based on these observations, we define a *specificity* constraint that states that only the most specific applicable concepts should appear in datasets. This constraint is typically satisfied in ecology and biodiversity datasets, and often appears in data collection best practices documents [14]. One ramification of this constraint is that it may not be possible to derive a single merged dataset that adheres to the specificity constraint. As shown in [15], it may be necessary to provide multiple possible data merges that adhere to the constraint. If, however, a user desires only one merged dataset, we will need to provide a *best-effort merge* which is as specific as possible.

C. Leveraging the Data Exchange Setting

Although there are some potential mismatches between our setting and that of traditional data exchange, we can still use the machinery of data exchange to help solve the problem of finding the best-effort merge. The following section sketches out this process.

II. Approach

The introduction described the best-effort merge of two taxonomically organized datasets. Several parts of the process for creating this merge need specification. These include (i) describing how a dataset is translated into a source instance, (ii) determining the source schema, (iii) determining the target schema constraints Σ_t, (iv) deriving the source-to-target dependencies Σ_{st}, and (v), describing how the target instance calculated using data exchange is translated into a dataset that satisfies the specificity constraint.

A. A Simple First Attempt

A first, straightforward translation is to take all concepts C in the input taxonomies $\mathsf{T}_1, \mathsf{T}_2$ and view them as a unary relations $\mathtt{C}(\mathtt{x})$ of the source schema $\mathbf{S}$, and consider concepts C' of the merged taxonomy T_3 as relations $\mathtt{C}'(\mathtt{x})$ of the target schema $\mathbf{T}$.[4] For presence/absence datasets in their most basic form, we can only state that a concept C is present or absent. We model this via facts of the form $\mathtt{C}(\mathtt{P})$ and $\mathtt{C}(\mathtt{N})$, representing that some instance of concept C was (or wasn't) observed.

1) Translating Taxonomies to Schemas: The next question is how to represent the taxonomy constraints. Using RCC-5, for any two non-empty sets A, B, exactly one of the following relations must hold: $\mathsf{A} \equiv \mathsf{B}$, $\mathsf{A} \subsetneq \mathsf{B}$, $\mathsf{A} \supsetneq \mathsf{B}$, $\mathsf{A} \oplus \mathsf{B}$, or $\mathsf{A} \,!\, \mathsf{B}$. Proper part ($\subsetneq$) and overlaps ($\oplus$) were described in the introduction. Identity ($\mathsf{A} \equiv \mathsf{B}$) is defined as $\forall x : \mathsf{A}(x) \leftrightarrow \mathsf{B}(x)$.

[4] Equivalent concepts in the input taxonomies $\mathsf{T}_1, \mathsf{T}_2$ are replaced by new concept names in T_3, denoting the equivalence class [6].

The last relation ($\mathsf{A} \,!\, \mathsf{B}$) refers to disjointness and can be represented as $\neg\exists x : \mathsf{A}(x) \wedge \mathsf{B}(x)$. Identity and proper part constraints can be approximated by the following constraints:

$$\mathsf{A} \equiv \mathsf{B} : \mathtt{A}(\mathtt{P}) \leftrightarrow \mathtt{B}(\mathtt{P}),\ \mathtt{A}(\mathtt{N}) \leftrightarrow \mathtt{B}(\mathtt{N})$$
$$\mathsf{A} \subsetneq \mathsf{B} : \mathtt{A}(\mathtt{P}) \rightarrow \mathtt{B}(\mathtt{P}),\ \mathtt{B}(\mathtt{N}) \rightarrow \mathtt{A}(\mathtt{N})$$

These formulas describe integrity constraints on a schema **S**. The first states that if $\mathsf{A} \equiv \mathsf{B}$ in the taxonomy, then $\mathtt{A}(\mathtt{P}) \in I$ iff $\mathtt{B}(\mathtt{P}) \in I$, and $\mathtt{A}(\mathtt{N}) \in I$ iff $\mathtt{B}(\mathtt{N}) \in I$. The second constraint states that if $\mathsf{A} \subsetneq \mathsf{B}$ in the taxonomy, then if taxon A has been reported as present in the dataset, then taxon B must also be present. In addition, if B has been reported as not present in the dataset, then A must also be reported as not present. Note that this is essentially an encoding in propositional logic, and that some information is lost in the process. For example, the disjointness relation does not constrain the schema in our case; even though A and B are disjoint, examples of each may be observed. Given this simple translation of concepts into relations, the partially overlaps relation does not translate into an integrity constraint.

2) Translating Datasets into Instances: Datasets rarely contain information about all the nodes in the accompanying taxonomy. However, given a source schema as translated above, a dataset may be used to populate much of the schema. For example, if $\mathsf{A} \subsetneq \mathsf{B}$, and $\mathsf{B} \subsetneq \mathsf{C}$, and the dataset contains (A, P), we can assert the presence of B and C. Similarly, if the dataset contains (C,N), we assert $\mathtt{A}(\mathtt{N})$ and $\mathtt{B}(\mathtt{N})$. This information can be used to determine inconsistent datasets. If a relation contains both P and N tuples, it is inconsistent.

3) Calculating Σ_{st}: In this simple scenario, Σ_{st} is given by the constraints above.

4) Translating Instances Back into Datasets: Once the data exchange setting is constructed, we can calculate a merged dataset by taking the source datasets, representing them as instances of the source schema, and following the dependencies to construct a target instance. To finish the process, that target instance should be translated back into a dataset.

As discussed in the introduction, we want to avoid ambiguity in our datasets. Unfortunately, the target instance as it stands will almost always be ambiguous. Even if there is only one dataset, with one observation "A is present", if $\mathsf{A} \subsetneq \mathsf{B}$, then the target instance calculated via data exchange will have $\mathtt{A}(\mathtt{P})$ and $\mathtt{B}(\mathtt{P})$. The $\mathtt{B}(\mathtt{P})$ is ambiguous because, once it is taken out of the context of the data exchange setting, it is unclear whether that $\mathtt{B}(\mathtt{P})$ is due to the $\mathtt{A}(\mathtt{P})$ or if there is some other B present which is not also an A.

One potential solution would be to only include the leaves of the merged taxonomy when translating from a target instance to a target dataset. However, in many cases this approach would lead to incorrect results. For example, consider two taxonomies of one concept each, A in T_1 and B in T_2. Dataset D_1, registered to T_1 reports the presence of A, dataset D_2, registered to T_2 reports the presence of B. The target instance contains $\mathtt{A}(\mathtt{P})$ and $\mathtt{B}(\mathtt{P})$, but converting these into a target dataset that only contains A would be incorrect. In particular, it may be the case that something that was a B but not an A was present.

5) Problems with the First Attempt: For data exchange to be useful, the target instance we generate either must adhere to the specificity constraint, or it must at least provide us with enough information to derive a dataset that adheres to the constraint.

There are many problems with the naive translation above, e.g., it is hard or impossible to enforce the specificity constraint when translating the target instance back into a dataset. The key difficulty in enforcing the constraint is that statements like "something that is a B but not also an A" cannot be represented if we restrict ourselves to discussing concepts from the taxonomies. Instead, to unambiguously describe something that is a B but not an A requires an enriched vocabulary.

B. Combined Concepts and Disjunctive Constraints

Aligning taxonomies provides additional information about the concepts in each taxonomy. For example, if concept A in one taxonomy overlaps with concept B in the second, we move from having two concepts (A and B) to having three (As that are Bs, As that are not Bs and Bs that are not As). Restricting our vocabulary to the original concepts A and B limits our ability to be specific about the outcome of a data merge. To overcome this limitation, we introduce the notion of *combined concepts*.

A combined concept is defined over the given taxonomy concepts using conjunction and negation. For example, two concepts A, B may be combined to create four new concepts, represented as AB, $\mathsf{A\bar{B}}$, $\mathsf{\bar{A}B}$ and $\mathsf{\bar{A}\bar{B}}$. Given certain taxonomic constraints, some combined concepts are unsatisfiable. If A and B are concepts of a taxonomy in which $\mathsf{A} \subsetneq \mathsf{B}$, the combined concept $\mathsf{A\bar{B}}$ can have no instances. Algorithms for computing the satisfiable combined concepts for a taxonomy are provided in [15]. Allowing the use of combined concepts in the target schema permits us to use the specificity added by the taxonomy alignments. For example, if A overlaps B, the combined concepts are AB, $\mathsf{A\bar{B}}$, $\mathsf{\bar{A}B}$, and the source to the target dependencies are:

$$\mathtt{A}(\mathtt{P}) \rightarrow \mathtt{A\bar{B}}(\mathtt{P}) \vee \mathtt{AB}(\mathtt{P})$$
$$\mathtt{A}(\mathtt{N}) \rightarrow \mathtt{A\bar{B}}(\mathtt{N}) \wedge \mathtt{AB}(\mathtt{N})$$
$$\mathtt{B}(\mathtt{P}) \rightarrow \mathtt{\bar{A}B}(\mathtt{P}) \vee \mathtt{AB}(\mathtt{P})$$
$$\mathtt{B}(\mathtt{N}) \rightarrow \mathtt{\bar{A}B}(\mathtt{N}) \wedge \mathtt{AB}(\mathtt{N})$$

These disjunctive dependencies result in a target instance that contains uncertainty [2]. We can query this target instance, e.g., to determine if each combined concept's presence or absence value is certainly known or not. Alternatively, we can materialize all possible instances. Tables I(a) and (b), for example, show two of the five possible merges[5] resulting from Figure 1. Each of the concepts described in Tables I(a) and (b) is a most specific applicable concept: it specifies whether or not it is subsumed by each of the original concepts. For example, the concept $\mathsf{A\bar{B}C\bar{D}}$ is subsumed by concepts A and C, but not subsumed by concept B or D.

[5] There are five possible merges under the assumptions that concepts A and B are disjoint and that there are no instances in concept C that are not in either A or B.

TABLE I
TWO POSSIBLE MERGES (A), (B) OF THE DATASETS IN FIGURE 1. A SINGLE BEST-EFFORT MERGE IS SHOWN IN (C).

Concept	***O***
$A\bar{B}C\bar{D}$	N
$A\bar{B}CD$	P
$\bar{A}BCD$	N
$\bar{A}\bar{B}CD$	N

(a)

Concept	***O***
$A\bar{B}C\bar{D}$	P
$A\bar{B}CD$	N
$\bar{A}BCD$	N
$\bar{A}\bar{B}CD$	P

(b)

Concept	***O***
C	P

(c)

C. Constructing the Most Specific Single World

The target instance follows a schema derived from the merged taxonomies. If the target instance contains uncertainty, (e.g., instances of the combined concept $\mathsf{A\bar{B}C\bar{D}}$ may or may not be present) we do not have a single most specific world. One way to get a most specific single world is to alter the merged taxonomy by removing concepts that contribute to the uncertainty. If uncertainty arises when internal (non-leaf) concepts are included in the target dataset, we can address this issue by removing all the leaves under the problematic internal concepts. This in effect makes the merged taxonomy less specific, but makes the dataset unambiguous in the context of the less specific taxonomy.

Table I(c) provides an example of a single best-effort merge for the scenario in Figure 1.

III. CONCLUSION AND SOME CHALLENGES

This paper provides a first, very rough sketch of how to leverage data exchange to merge datasets that draw their concepts from taxonomies. Fleshing the process out more completely and giving it a formal foundation is the first priority in future work on the problem. Assuming that the process described here can be used to create a best-effort merge, further issues arise:

A. Disjunctive Relations

In this paper we have restricted our articulation relations to the basic five set relations. RCC-5 also encodes disjunctive relations, e.g., $(\mathsf{A} \subsetneq \mathsf{B}) \vee (\mathsf{A} \equiv \mathsf{B})$ is the standard definition of "isa". There are 32 disjunctions over the basic five relations.

B. Restricting the Target Instance to Original Concept Terms

An effect of our move to combined concepts is that the resulting merged dataset will most likely use a vocabulary different from the vocabulary used in the original datasets. This calls for a second translation from the calculated dataset to one using only terms from the original taxonomies.

C. Translation from the Target Instance to a Dataset

It remains to be seen how to derive from the target instance a merged dataset that adheres to the specificity constraint, and how to encode the constraint as dependencies on the target instance.

ACKNOWLEDGMENT

This work was supported in part by NSF awards IIS-0630033 (pPOD), DBI-0743429 (Semtools), and DBI-0753144 (SONet).

REFERENCES

[1] R. Fagin, P. G. Kolaitis, and L. Popa, "Data exchange: getting to the core." in *PODS*. ACM, 2003, pp. 90–101.

[2] R. Fagin, P. G. Kolaitis, L. Popa, and W. C. Tan, "Quasi-inverses of schema mappings." in *PODS*, 2007, pp. 123–132.

[3] M. Arenas and L. Libkin, "XML data exchange: Consistency and query answering." *Journal of the ACM*, 55(2), pp. 1–72, 2008.

[4] D. A. Randell, Z. Cui, and A. Cohn, "A spatial logic based on regions and connection," in *KR'92*, B. Nebel, C. Rich, and W. Swartout, Eds. San Mateo, California: Morgan Kaufmann, 1992, pp. 165–176.

[5] D. Thau and B. Ludäscher, "Reasoning about taxonomies in first-order logic," *Ecological Informatics*, 2(3), pp. 195–209, 2007.

[6] D. Thau, S. Bowers, and B. Ludäscher, "Merging taxonomies under RCC-5 algebraic articulations." in *ONISW*, R. Elmasri, M. Doerr, M. Brochhausen, and H. Han, Eds. ACM, 2008, pp. 47–54.

[7] K. K. Iguchi, K. Matsuura, A. McNyset, *et al.*, "Predicting invasions of North American basses in Japan using native range data and a genetic algorithm," *Transactions of the American Fisheries Society*, vol. 133, pp. 845–854, 2004.

[8] A. T. Peterson, D. A. Vieglais, and J. K. Andreasen, "Migratory birds modeled as critical transport agents for West Nile virus in North America," *Vector-Borne and Zoonotic Diseases*, 3(1), pp. 27–37, 2003.

[9] D. B. Botkin, H. Saxe, *et al.*, "Forecasting the effects of global warming on biodiversity." *BioScience*, vol. 57, pp. 227–236, 2007.

[10] J. Kennedy, R. Kukla, and T. Paterson, "Scientific names are ambiguous as identifiers for biological taxa: Their context and definition are required for accurate data integration," in *Intl. Workshop on Data Integration in the Life Sciences (DILS)*, LNCS 3615, July 2005, pp. 80–95.

[11] N. M. Franz, "Taxonomic concept mappings for select higher-level classifications of weevils (Coleoptera: Curculionoidea) published from 1981 to 2000." 2005, unpublished dataset.

[12] M. Koperski, M. Sauer, W. Braun, and S. Gradstein, *Referenzliste der Moose Deutschlands*. Schriftenreihe Vegetationskunde, 2000, vol. 34.

[13] R. K. Peet, "Taxonomic concept mappings for 9 taxonomies of the genus Ranunculus published from 1948 to 2004." June 2005.

[14] A. D. Chapman, "Principles of data quality," Global Biodiversity Information Facility, Copenhagen, Tech. Rep., 2005.

[15] D. Thau, S. Bowers, and B. Ludäscher, "Merging sets of taxonomically organized data using concept mappings under uncertainty." in *OTM Conferences*, LNCS 5871. Springer, 2009, pp. 1103–1120.