Provenance for Explaining Taxonomy Alignments

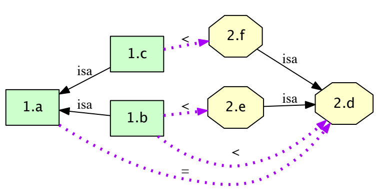
Mingmin Chen[1], Shizhuo Yu[1], Parisa Kianmajd[1], Nico Franz[2], Shawn Bowers[3], and Bertram Ludäscher[1]

[1] Dept. of Computer Science, UC Davis, {michen,szyu,pkianmajd,ludaesch}@ucdavis.edu
[2] School of Life Sciences, Arizona State University, nico.franz@asu.edu
[3] Dept. of Computer Science, Gonzaga University, bowers@gonzaga.edu

Derivations and proofs are a form of provenance in automated deduction that can assist users in understanding how reasoners derive logical consequences from premises. However, system-generated proofs are often overly complex or detailed, and making sense of them is non-trivial. Conversely, without any form of provenance, it is just as hard to know why a certain fact was derived.

We study provenance in the application of Euler/X [1], a logic-based toolkit for aligning multiple biological taxonomies. We propose a combination of approaches to explain both, logical inconsistencies in the input alignment, and the derivation of new facts in the output taxonomies.

Taxonomy Alignment. Given taxonomies T_1, T_2 and a set of *articulations* A, all modeled as monadic, first-order constraints, the *taxonomy alignment problem* is to find "merged" taxonomies that satisfy $\Phi = T_1 \cup T_2 \cup A$. An alignment can be *inconsistent* (Φ is unsatisfiable), *unique* (Φ has exactly one minimal model), or *ambiguous* (Φ has more than one minimal model). For example, let T_1 be given by *isa* (subset) constraints $\mathsf{b} \subseteq \mathsf{a}$, $\mathsf{c} \subseteq \mathsf{a}$, *coverage* constraint $\mathsf{a} = \mathsf{b} \cup \mathsf{c}$, and *sibling disjointness* $\mathsf{b} \cap \mathsf{c} = \emptyset$. Similarly, T_2 is given by isa constraints $\mathsf{e} \subseteq \mathsf{d}$, $\mathsf{f} \subseteq \mathsf{d}$, coverage $\mathsf{d} = \mathsf{e} \cup \mathsf{f}$, and sibling disjointness $\mathsf{e} \cap \mathsf{f} = \emptyset$.



Taxonomy T_1	Taxonomy T_2	Articulations
r_1: b isa a	r_5: e isa d	A_1: $\mathsf{a} = \mathsf{d}$
r_2: c isa a	r_6: f isa d	A_2: $\mathsf{b} \subsetneq \mathsf{e}$
r_3: $\mathsf{a} = b \cup c$	r_7: $\mathsf{d} = e \cup f$	A_3: $\mathsf{c} \subsetneq \mathsf{f}$
r_4: $\mathsf{b} \cap \mathsf{c} = \emptyset$	r_8: $\mathsf{e} \cap f = \emptyset$	A_4: $\mathsf{b} \subsetneq \mathsf{d}$

Fig. 1. Alignment Problem: Taxonomies T_1 (given by set constraints $r_1, \ldots, r_4$) and T_2 (constraints $r_5, \ldots, r_8$) are related via articulations A (constraints $A_1, \ldots, A_4$).

An expert aligns T_1 and T_2 using *articulations* $\mathsf{a} = \mathsf{d}$, $\mathsf{b} \subsetneq \mathsf{e}$, $\mathsf{c} \subsetneq \mathsf{f}$, and $\mathsf{b} \subsetneq \mathsf{d}$; see Figure 1. We would like to "apply" all of these relations between the two taxonomies, and output a merged taxonomy.

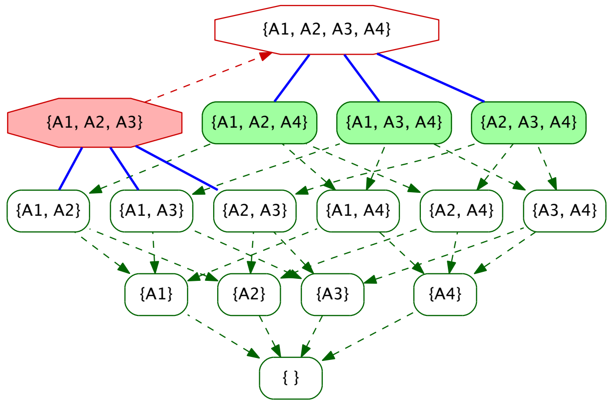


Fig. 2. Diagnosis for $A = \{\mathsf{A}_1, \ldots, \mathsf{A}_4\}$: solid red octagons and solid green boxes denote MIS and MCS, respectively. The (in)consistency of all other combinations are implied.

Inconsistency Explanation. Usually T_1 and T_2 are considered immutable or correct by definition, whereas A might contain modeling errors. Euler/X applied to Fig. 1 finds that the constraints are unsatisfiable, and performs a model-based diagnosis. The result lattice (Fig. 2) highlights *minimal inconsistent subsets* (MIS) and *maximal consistent subsets* (MCS). The MIS $\{\mathsf{A}_1, \mathsf{A}_2, \mathsf{A}_3\}$ indicates which articulations are inconsistent with T_1, T_2. To further explore the inconsistency, the system-derived MCS can be employed: Fig. 3 shows the merged taxonomies (a.k.a. "possible worlds") obtained from the MCS. Here, each MCS corresponds to one possible world.[4]

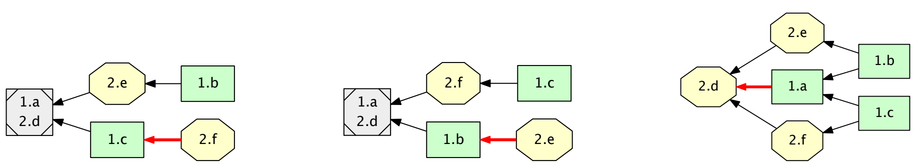


Fig. 3. Merged taxonomies (*possible worlds*) for MCS $\{\mathsf{A}_1, \mathsf{A}_2, \mathsf{A}_4\}$, $\{\mathsf{A}_1, \mathsf{A}_3, \mathsf{A}_4\}$, and $\{\mathsf{A}_2, \mathsf{A}_3, \mathsf{A}_4\}$. Grey boxes are fused concepts; bold, red edges represent inferred relations.

Using expert knowledge or further constraints[5] a preferred merge result can be selected to further analyze and then repair the inconsistency. Here, suppose the user chose the first maximal consistent subset $\{\mathsf{A}_1, \mathsf{A}_2, \mathsf{A}_4\}$. It follows from $\mathsf{A}_1, \mathsf{A}_2$ and the input taxonomies T_1, T_2 that $\mathsf{f} \subsetneq \mathsf{c}$. However, A_3 is $\mathsf{c} \subsetneq \mathsf{f}$ yielding a contradiction. Now the problem is to explain why $\mathsf{f} \subsetneq \mathsf{c}$ is inferred.

[4] In general, a MCS can yield many possible worlds. Such ambiguities arise when the alignment input is underspecified.

[5] E.g., the output for MCS $\{\mathsf{A}_2, \mathsf{A}_3, \mathsf{A}_4\}$ might be less desirable since it is not a tree.

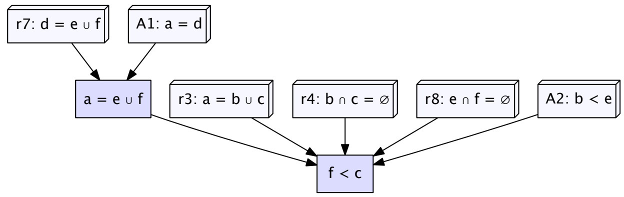


Fig. 4. Provenance of $\mathsf{f} \subsetneq \mathsf{c}$ (depicted as $\mathsf{f}{<}\mathsf{c}$). Lightly colored 3D-boxes are input facts (taxonomies and input alignment). Inferred relations are shown as darker boxes.

Derivation Explanation. To understand how $\mathsf{f} \subsetneq \mathsf{c}$ is inferred, we may need to inspect its logical derivation or an abstraction of it. We obtain this provenance in Euler/X by keeping track of the rules $r_1, \ldots, r_8$ and input alignments $\mathsf{A}_1, \ldots, \mathsf{A}_4$ used by the reasoner. Fig. 4 depicts the resulting provenance overview.

Related Work. Data provenance is an actively researched area and is closely related to proofs and derivations in logical reasoning. Our inconsistency explanation is based on Reiter's model-based diagnosis [6], which has been studied extensively and applied to many areas, e.g., type error debugging, circuit diagnosis, OWL debugging, etc. We have adapted the HST algorithm in [4] to compute all MIS and MCS for inconsistency explanation. The problem was shown to be Trans-Enum-complete by Eiter and Gottlob [2]. Inspired by the ideas of a provenance semirings [3] and Datalog debugging [5], our approach explains the derivation of the inferred relations.

Acknowledgments. Supported in part by NSF IIS-1118088 and DBI-1147273.

References

1. M. Chen, S. Yu, N. Franz, S. Bowers, and B. Ludäscher. Euler/X: A toolkit for logic-based taxonomy integration. In *22nd Intl. Workshop on Functional and (Constraint) Logic Programming (WFLP)*, Kiel, Germany, 2013.
2. T. Eiter and G. Gottlob. Hypergraph transversal computation and related problems in logic and AI. In *European Conference on Logics in Artificial Intelligence (JELIA)*. LNCS 2424, Springer, 2002.
3. T. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 31–40, 2007.
4. M. Horridge, B. Parsia, and U. Sattler. Explaining inconsistencies in OWL ontologies. In *Scalable Uncertainty Management*, LNCS 5785, Springer, 2009.
5. S. Köhler, B. Ludäscher, and Y. Smaragdakis. Declarative datalog debugging for mere mortals. In *Datalog in Academia and Industry*, LNCS 7494, Springer, 2012.
6. R. Reiter. A theory of diagnosis from first principles. *Artificial intelligence*, 32(1):57–95, 1987.