A Superstructure for Organizing Family History Information

David W. Embley and Scott N. Woodfield

Abstract

To make computers work better for us in the family history domain, they must automatically process certainty and conflicting information, support evidence-based research, automate collaboration, and provide research guidance. To address these issues, we propose a superstructure that adds four additional abstraction layers to typical conceptual models: the knowledge, evidence, communication, and action layers. We show, using a running example, the benefits these abstraction layers provide for organizing and processing family history information.

1 Introduction

Five significant problems that reduce our capacity to more meaningfully automate family history research are:

- Our inability to record and automatically process certainty information.
- Our inability to record and automatically process conflicting information.
- The lack of support for doing evidence-based family history research.
- The difficulty we have in automating collaboration effectively.
- The challenge of developing automated research assistants to guide us in our efforts.

A primary cause of all these problems is the implicit use of weak underlying conceptual models to organize information. Decades ago programmers coded in assembly language, and the information structure used to organize data resided primarily in a developer's head. In the 1960s and 70s, higher-level languages and their libraries supported explicit organizing concepts such as sets, sequences, aggregates, trees, and maps. On the heels of high-level language usage, we began to work with databases to organize information, and the relational model came into common use. Structuring data with even higher-level abstractions was spurred on by the conceptual modeling community, but beyond its use to aid in generating database schemas, it has not had much impact in organizing data. Since the advent of relational databases, little has changed in our quest to organize and process data at higher levels of abstraction. While necessary, these structures are not sufficiently powerful to enable us to easily understand and solve the five problems hindering better use of automation in solving family history research problems. We need more powerful data-organization and data-processing tools.

To understand and begin to solve the enumerated problems we propose a more powerful conceptual framework for information organization—a 7-layered superstructure. With it we can decompose the overall problem into decoupled pieces that hide lower-level detail and support work at an abstraction layer closer to human-level genealogical research. Our 7-layer organization is not without precedent, as we base it on a solid theoretical foundation suggested by Charles T. Meadow.[1]

- Symbol Layer: to represent the atomic symbols or tokens of recorded information.
- Class Layer: to classify and provide semantics for symbols.
- Information Layer: to describe relations between classes and constraints over their instances.
- Knowledge Layer: to allow for conflicting information and informal assertion verification.

[1] C.T. Meadow, *Text Information Retrieval Systems*, Academic Press, 1992.

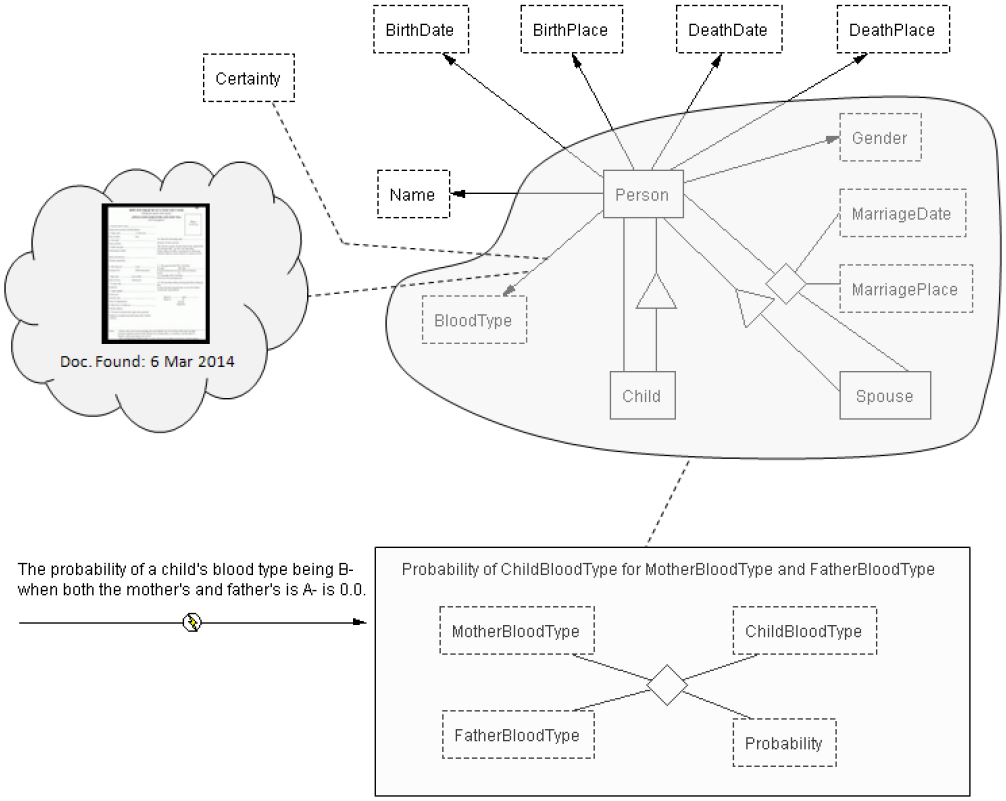

Figure 1: Symbol/Class/Information/Knowledge/Evidence/Communication Conceptualization.

- Evidence Layer: to organize and record evidence to support and automate evidentiary logic.
- Communication Layer: to send and receive structured information without distortion or loss.
- Action Layer: to allow automated expert mentors to guide user behavior.

2 Conceptualization Superstructure

Figure 1 and 2 are conceptual-model diagrams, which we use to illustrate our superstructure. In the subsections below, we explain the components of these diagrams that pertain to each of the seven layers and how these components provide an increase in power over each preceding layer. As an illustration of the superstructure, we show how the increase in power helps solve a family-history research problem—discovering the parentage of Stan Williams, initially asserted to be the son of Roger and Judith Williams, but refuted based on blood-type information.

2.1 Symbol Layer

Conceptualization: The symbol layer has no conceptual-modeling features. It is a collection of symbols organized as text files or image documents, which we represent by the cloud in Figure 1.

Example: We may have a copy of Judith Williams' diary. Entries in the diary pertinent to our story tell about raising Stan and also mentioning that the results of a blood-type test for her and her husband are both "A-".

Superstructure Motivation: Text documents and images are easily understood by humans, but difficult for machines to organize or process. With no conceptualization facilities, the symbol "A-" appearing in the diary has no semantic meaning that a computer can rely on for processing.

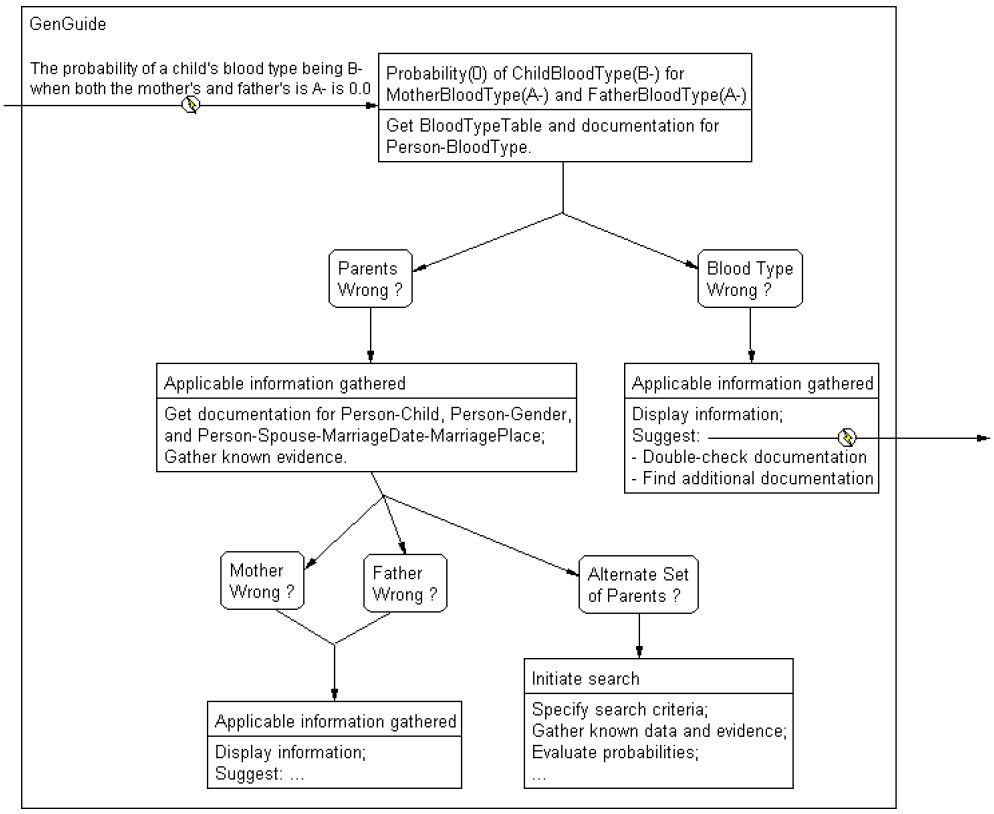

Figure 2: Action Conceptualization.

2.2 Class Layer

Conceptualization: For given symbols we are able to state, especially for nouns, what class of thing they represent. In Figure 1, named rectangles represent classified sets of objects.

Example: Recording the symbol "A-" in the *BloodType* object set in Figure 1 signifies that it is a blood type and that the computer may therefore operationally manipulate it as such.

Superstructure Motivation: The computer can't record how this value relates to anything else. With classes only, we cannot associate a *BloodType* with a *Person*.

2.3 Information Layer

Conceptualization: All typical conceptual modeling features are included at this level: classes, relationships, generalization/specialization, and cardinality constraints. In Figure 1, lines connecting object sets represent relationship sets (e.g., *Person-BloodType*). Lines with triangles represent *isa* abstractions (e.g., *Child-isa-Person*). Decorations on lines express cardinality constraints (e.g., an arrowhead designates a functional constraint from tail object set to head object set).

Example: We can now record the blood type "A-" for Stan, assumed so, since his parents are both "A-".

Superstructure Motivation: At the information layer, models must be valid—all constraints must hold—but in family history research we often have conflicting information. Furthermore, we would like to be able to justify assertions and reason about assertions with soft constraints such as distributions, which are not enforceable in the same way as hard constraints.

2.4 Knowledge Layer

Conceptualization: This layer allows for invalid models, soft constraints, and unstructured meta-information for justification. In Figure 1 the dashed line associates the unstructured information in the cloud with the relations in the *Person-BloodType* relationship set.

Example: A discovered military form states that Stan has blood type B-. Since we can violate constraints we can add this assertion along with the assertion that his blood type is "A-". Further, we can attach the military form as justification, as Figure 1 shows. We can also give a soft constraint about blood type distribution and use it as justification for Stan's "A-" blood type (96.75% chance since both parents are "A-").

Superstructure Motivation: Although we can record evidence for assertions, the computer cannot use the information to reason, because the information is informal.

2.5 Evidence Layer

Conceptualization: This layer allows the evidence to be formally organized in a conceptual model. In Figure 1 the 4-ary relationship *Probability of ChildBloodType for MotherBloodType and FatherBloodType* formally gives a table of blood type probabilities given parents' blood types.

Example: Because the information is formal, the computer can reason about Stan's blood type and discover that if Stan's blood type is "B-" and Roger and Judith's are both "A-", then there is a 0% chance Stan is their child.

2.6 Communication Layer

Conceptualization: To communicate on its own, the computer must be able to read and write and to send and receive information. Figure 1 shows a message coming into a conceptualization about blood type.

Example: In our running example, we imagine that upon discovery of the military record, Judith's descendant asks a doctor friend about it and receives the message in Figure 1 in reply. If the model includes extraction ontologies[2] as part of its conceptual model, it can automatically read the message, populate itself with the information, and reason that something is wrong.

2.7 Action Layer

Conceptualization: At this level of abstraction, we add object-behavior modeling along with object-interaction modeling and object-relationship modeling.[3] Figure 2 shows an example.

Example: Knowing that something is wrong, the computer can direct Judith's descendant in further research as Figure 2 shows. Either the military record is wrong about Stan's blood type, or Roger and Judith are not his parents.

2.8 Conclusion

Meadow adds a last layer which he calls "Wisdom", and which we interpret to mean: the proper application of knowledge based on truth (via evidence), communication, and action. Because the 7-layer superstructure lets us record and process richer information, perform evidence-based reasoning, collaborate effectively, it can more effectively enable us to proceed wisely and further help us "turn ... the heart of the children to their fathers".[4]

---

[2] D.W. Embley, et al., "Conceptual-model-based data extraction from multiple-record web pages", *Data & Knowledge Engineering*, 31(3), November 1999, 227–251.

[3] D.W. Embley, B.D. Kurtz and S.N. Woodfield, *Object-oriented Systems Analysis: A Model-driven Approach*, Prentice-Hall, Inc., 1992.

[4] Malachi 4:6