Towards Ontologies in Variation (Extended Abstract)

Torsten Hahmann
School of Computing and Information Science
University of Maine, Orono, ME 04469, USA
torsten@spatial.maine.edu

Sheila A. McIlraith
Department of Computer Science
University of Toronto, Ontario, Canada M5S 3G8
sheila@cs.toronto.edu

1 Introduction

In a small laboratory within the Anatomy Department at the University of Toronto, researchers are delicately pealing muscles off of cadavers, strand by strand. While this may sound like the type of macabre activity reserved for oddball television crime shows, these scientists are working diligently in an effort to discern and record measurable difference in human bodies. The work is part of a larger project called the Parametric Human Project (Mogk et al. 2013) whose goal is to create a parametrized digital model of the human body, ultimately in all its variations, in support of such tasks as computer animation and surgery planning. Within this ambitious project is an intriguing knowledge representation and reasoning (KR) challenge: How to represent our knowledge of human anatomy in a human- and computer-interpretable form – a sort of queryable digital anatomy book, in the spirit of the classic Gray's Anatomy textbook. However, whereas such books capture the prototypical (or "canonical") human – an idealized human that, in reality, is like very few if any of us – our objective is to create a more inclusive representation of the human anatomy that fits *all* of us in some variation.

While more flexible representations of human anatomy motivate our research, our more general interest is in examining the principles that underlie the construction of what we call *ontologies in variation* – a KR scheme for natural kinds, objects, concepts – *Things*, for lack of a better word – in all their variation. Not unlike a typical machine learning classifier, our assumption is that this variation is derived from statistical data and that this statistical information is preserved in the representation. Nevertheless, whereas machine learning classifiers may represent an object in terms of a set of low-level features, aggregated together arithmetically, often with little inherent meaning to a human examining this representation, what is unique about the class of ontologies we explore is that the features we use to characterize Things are themselves Things of meaning to humans. Further, the method of association or composition of these features is intended to be in terms of human-understood, measurable *properties* of these Things and the *relations* between them.

To realize our ontologies in variation, we appeal both to first-order logic (FOL) based KR and to statistical representation of Things from data. While FOL-based KR schemes are highly expressive along many dimensions, they are impoverished in their ability to succinctly capture nuanced variations in the characterization of concepts, relative to statistical representations. Indeed, logic-based KR typically captures necessary, but rarely sufficient, conditions that hold true for a concept. Variability is largely captured using disjunction or through the creation of named subclasses of concepts. The notion of typicality is often only achieved via nonmonotonic reasoning. When logical languages are used to describe what can be said about concepts described natively in terms of statistical data, the characterizations are necessarily weak and afford little inferential power.

There has been significant interest in the merging of logical and statistical or probabilistic information, e.g., in (Bacchus 1990; Domingos et al. 2006; Kuo and Poole 2013; Poole et al. 2008). A distinguishing feature of our work is an explicit focus on data, and a frequentist interpretation of statistical information. This paper outlines our endeavour and approach, and contrasts it to existing work in the literature.

2 Use Case: Characterizing Human Bones

The desire to create a representation for capturing and understanding typical and variable morphology in human musculoskeletal anatomy (Hahmann et al. 2014; Mogk et al. 2013) motivates and illustrates our research. Building a digital representation of anatomy is not an entirely new endeavour, e.g., the Foundational Model of Anatomy (FMA) ontology (Rosse and Mejino Jr. 2003) has accumulated several thousand classes and over 2 million relationship instances. However, the FMA represents a single *canonical human* deemed free of pathologies – a very abstract notion that is of little help for understanding human variation and for understanding how an individual person compares to a larger populace. Moreover, due to the ontology's unmanageably large size and its manual curation, it inevitably contains factual errors/inconsistencies and misses crucial information.

Other approaches (e.g., Schulz and Hahn 2007), capture anatomical variability by dividing anatomical knowledge into levels: one including only the axioms that all humans (with or without pathologies) satisfy, while the next one adds axioms that applicable only to the canonical human. Such a purely symbolic representation faces the same obstacles as all logic-based KR approaches: it is ill-suited to capture nuances of variability. Moreover, it makes the strong assumptions that (1) an a-priori definition of what the canonical conditions are, and that (2) all of these conditions can be captured manually in a purely symbolic description. We put these assumptions to test: What is really canonical about human anatomy, and what portion of this knowledge can be described axiomatically, and what is better described using automatic statistical aggregations of data?

As an initial step (Hahmann et al. 2014), we tested our ideas on characterizing classes of bones (starting with tibias) by identifying, relating, and quantifying anatomical landmarks on the bones' geometric surfaces, extracting invariants and variability of each class of bones. We relied on an ontology based on the Terminologia Anatomica (TA) (FIPAT 2011), which standardizes anatomical nomenclature internationally and which overlaps significantly with the FMA's taxonomy of classes. The TA-based ontology names the classes of bones we wanted to characterize, such as *tibia*, *fibula*, *sternum* and more general classifications such as *long*, *short*, *flat*, or *irregular bone* (generally: the classes of Things), and the anatomical landmarks (also Things), properties, and relations we would rely on for this characterization. The anatomical terms are complemented by spatial terms from (Hahmann 2013), which are largely lacking in the FMA, including names for classes (e.g. *surface areas*, *ridge lines*), relations (e.g. *overlaps*, *on the boundary of*), and properties (e.g. *volume*, *length*, *distance*).

We wanted to determine the typicality and variation within each class of bones, eventually answering the following kinds of questions: What do all tibias have in common – which anatomical landmarks are always present, which spatial relations always hold between them hold, and what are lower/upper limits for the landmarks' measurements (such as a ridge's minimal length or the minimal ratio between two named surface areas)? What are the modes of variations (the pathologies)? What does a typical tibia look like? We also wanted to be able to classify new Things, to complete an incomplete description of a Thing with suitable missing Things/relations/properties, to extract a set of Things that satisfy certain conditions, or to compare a subpopulation's variability (defined by a narrower reference class, e.g., all Inuit Females of age 20-29) to that of a broader population.

This use case has evoked a number of questions useful for evaluating whether our proposed solution adequately represents typicality and variability in human anatomy: What schemes of axiomatic knowledge[1] can be identified? Does the inferred axiomatic knowledge correspond to the textual descriptions of bone classes in the TA? What kind of medical/anatomical queries are expressible in the language? Are their answers easily interpretable by domain experts? Can we computationally verify and refine descriptions like "The shape of the shaft most frequent in both the white male and the female is that of a prism. About three-fifths of all tibiæ are of this variety, [...]" (Hrdlicka 1898) used to summarize manual examinations of 2,000 tibias?

Answers to these domain-specific questions will indicate whether the proposed solution can potentially address the much more general problem of capturing typicality and variability of Things, no matter the domain.

3 Approach

The beauty of FOL-based knowledge representation is that – by design – the nonlogical language, i.e., the vocabulary used to talk about a domain, closely resembles the terms that humans use when talking about that domain. Thus, any logical inferences are generally human-comprehensible and humans have explicit access to the representation, allowing them to effect changes in the representation by altering, adding, or deleting axioms. The caveat is that it is extremely difficult to express all nuances of variability within classes in a logical representation. To capture variability, some statistical representation is required. To benefit from logic-based and statistical representations, we are naturally inclined to combine them into a single more powerful representation. The question is *how* exactly to combine them?

Our proposed formalism is guided by the various ways in which we envision it being utilized: It is not only designed to find applications in a variety of knowledge domains but is also meant to support a broad range of different kinds of inference/retrieval tasks, for example (**t** denotes a particular Thing and X and Y particular classes):

1. *Object Classification Tasks:* What is the most suitable class for **t**? How likely does **t** fit class X? Is it consistent to classify **t** as being in X? What additional information is needed to reliably classify **t**? What additional data would be most useful to confirm or rule out a classification for **t**? What information prevents **t** be classified as X?
2. *Purely Logical Reasoning:* Is X a subclass of Y? Can Things in X be related by relation r to Things in Y? May Things in X have a certain property p?
3. *Statistical Querying:* What are the mean/min/max values for property p of Things in X? How many (mean/min/max) relations r have Things in X to Things in Y?
4. *Object Retrieval Tasks:* What are all the instances of X? What are all the instances of X that have property p in a specific range (absolute or relative to its statistical information)? What is a representative sample of X? What are extreme cases of X? What Things are most similar to **t**?
5. *Inductive Reasoning:* What aspects of **t** are likely to be "abnormal" given what we know about **t**? How do Things in X differ from other Things in its superclass? Given what we know about **t**, what other properties and relations is it most likely to have (object completion)?
6. *Clustering:* Are there meaningful ways to divide the Things in X into two or more disjoint subclasses? What properties/relations would differentiate the clusters?

In addition to human-comprehensibility, our desiderata include logical consistency, and elaboration tolerance. Defining the former is more complex than pure logical consistency, while achieving the latter may be elusive.

[1] We use the term *axiomatic knowledge* to distinguish knowledge that is definite – over which there is no designated uncertainty. In our use case these are conditions satisfied by all humans.

4 Outline of our Formalism

Broadly speaking, we augment FOL with statistical functions ranging over special terms (*data collection terms*) that denote finite collections of Things (class instances), relation instances, or property instances rather than individuals. Applying statistical functions to such collections allows us to make statistical claims about them. For example, we may define a data collection consisting of all Things classified as tibias and having a convex posterior surface.

Our starting point is (a) an ontology that provides the shared vocabulary for describing the classes of interest as well the relations between them and their properties and (b) a dataset containing raw datapoints.

The **ontology** captures a-priori knowledge about the taxonomic relationships between the classes of Things in the domain and about the relations and properties the members of each class can participate in, formalized either in FOL or in a more restricted language such as OWL, furnishing *at least* the following information.

- a set of class names $\mathbf{C}$ and taxonomic relations between them that specify subclasses and disjointness and exhaustiveness conditions. $\mathit{Thing} \in \mathbf{C}$ denotes the universal class of which all Things are a member of;
- a set of n-ary (with $n \geq 2$) relation names $\mathbf{R}$ and associated typing constraints, which state which classes' members can participate in a specific place in a relation;
- a set of n-ary property names $\mathbf{P}$ and associated typing expressions (as for relations) for the places 1 to $n-1$ of each property, as well as a typing of the kind of acceptable measurement values (place n) for each property;

The ontology may include arbitrary additional axioms. Relations will hold between instances of the ontology's classes (classes of Things), while properties assign numeric, Boolean, or labeled values to instances of Things. In our use case, relations are primarily of mereonomic (part-of) or spatial nature, including the *partOf*, *surfaceOf*, *bounds*, and *attachedTo* relations, while properties are either quantitative measurements (assigning a numeric value to, e.g., *length*, *volume*, *curvature*, or *distanceBetween*) or qualitative categories (assigning labels, e.g., *conical*, *cylindrical*, or *prismatic* to the property *principal shape* or Boolean values to the properties *flat* or *closed*). The ontology forms a *logical shell* for the domain, elaborated by the datasets (microdata) and their statistical aggregations (macrodata).

Datapoints are facts about particular Things – similar to tuples in a database table. Logically, each datapoint is an atomic term of the form $X(c_1, \ldots, c_n)$ where X is a predicate – either a unary predicate denoting a class name or a predicate with higher arity (denoting a relation or property name) – and c_1 to c_n are constants. We distinguish *object constants*, each denoting one Thing, from *measurement constants* denoting the values that properties can take on. The standard measurement constants include numbers (integers, reals), Boolean values, or sets of labels. All parameters in a data point must be object constants, except if X is a property name, then c_n must denote a measurement constant. No variables may occur in datapoints.

Datasets are collections of *datapoints* collected under controlled conditions[2], such as all the data collected within a single experiment or study. This ensures that the individual datapoints in the data set are comparable and thus amendable to meaningful statistical aggregation. For simplicity we assume datasets to be noise-free and sufficiently large to contain a representative sample of Things of each class.

Data collection terms specify subcollections of Things, relation instances (a relation collection), or property instances (a property collection) from a dataset. Each data collection term only allows specifying subcollections of a single dataset so that incomparable datapoints are never mixed up[3]. We can construct data collections by naming individual Things (using object constants) or by referring to named classes of Things, as well as through intersections, unions, and differences of such collections. Collections can also be restricted to certain values of a property.

While we can determine the size of any collection, only the property values of property collections can be evaluated statistically. For example, a collection of the "length" values of all tibias in a dataset can be statistically evaluated with respect to, e.g., the min/mean/max length value. But to statistically evaluate a relation collection, it must be first converted into a property collection using a distinguished *aggregate* function, which operates on a relation or property collection and an associated list of the parameters that will be kept distinct (the *select* part of a *group-by* construct in SQL). The aggregate function results in a property collection with the property value being a count of the distinct property/relation instances that are grouped together.

Applying **statistical functions** such as median, mean, min, max, or standard deviation to a property collection forms a **statistical term**, which denotes again an individual measurement constant (e.g., an average or minimal value). Thus, statistical terms can be embedded into more complex FOL formulas like any other ordinary FOL terms. Then, we can logically express the statement "bone X's length is smaller than the mean length of tibias in dataset Y" or that "the mean length of tibias is smaller in dataset Y (for Inuit Females) than in dataset Z (for Caucasian Females). These statistical terms describe variability of a class, though some of them (such as those about the min/max number of relations that members of a particular class participate in, such as "every human has at least one heart" or min/max values of properties) impose strict logical conditions – definite constraints – that describe the canonical human.

5 Related Work

There is a large body of work that explores various synergies between logic-based and statistical/probabilistic KR methods (Demey et al. 2009), which we do not do justice to in this paper. While, superficially, a number of formalisms appear suitable to our endeavour, we discuss how they deviate from our objectives and/or approach.

[2] Datasets here are similar to those of (Poole et al. 2008).

[3] The language still supports logical statements that statistically compare, e.g., a property's mean value, between datasets.

A fundamental characteristic of our work is that the characterizations of variability are derived from data and as such the provenance of that statistical knowledge – the datasets – is directly associated with the ontology. This reliance on empirical data directs us towards a frequentist view of probabilities as statistical assertions about proportion or relative frequency, as opposed to the popular interpretation of probabilities as degrees of belief seen in most of the related work (see, e.g., (Halpern 1990; Bacchus et al. 1996) for a discussion of related issues.) In this regard, Bacchus' work on a logic for representing and reasoning with statistical knowledge (Bacchus 1990) shares commonalities with our work, although a key contribution of that work was the representation of qualitative statistical knowledge. Further, Bacchus makes no explicit connection with empirical data.

Most approaches to integrating logic and statistical/probabilistic knowledge rely on adding a probabilistic semantics to a classical logic by assigning each sentence in the logic's object language a probability value or subinterval in the interval $[0, 1]$ instead of a Boolean truth value. Moreover, such approaches rely exclusively on probabilistic inference even when dealing with only axiomatic knowledge. Rather, we restrict the kind of statistical information we support in favour of a concise representation and more efficient reasoning. While we severely restrict reasoning with statistical knowledge, we do so in ways that preserve statistical information relevant to a concept's variability.

One of the most popular recent formalisms that integrates logic and probability is Markov Logic Networks (MLN) (e.g. Domingos et al. 2006), a probabilistic logic that incorporates Markov Networks into FOL to support reasoning under uncertainty. In MLN, FOL formulae are assigned a real-valued weight that indicates the certainty of a formula; axiomatic knowledge is treated as a special case of probabilistic knowledge. MLN is a powerful framework for unifying axiomatic and probabilistic knowledge, but because of its generality, the conciseness of logic is lost and reasoning/querying is often more difficult than necessary. Moreover, we are unaware of proposals that ground MLN in real data. Similar to MLN, various probabilistic extensions to description logics have been proposed (Predoiu and Stuckenschmidt 2010). All of these extensions assign numeric probabilities or probability intervals to logical sentences, but do not explicitly distinguish between axiomatic (i.e., "definite") sentences and sentences with some uncertainty.

Poole and collaborators' work on KR for Relational Semantic Science (e.g., Kuo and Poole 2013; Poole et al. 2008) is similar to our work in two regards: (1) it extends an ontology with a statistical/probabilistic representation and with data, and (2) it uses the ontology as the source for the classes, properties, and relations of interest. Nevertheless, Poole's focus is on probabilistic inference for inductively testing how likely a hypothesis is to be true for a given set of observations. In contrast, our work is concerned with developing hybrid characterizations of classes for a broader range of applications, which include statistical queries as well as probabilistic reasoning (cf. Sec. 2). Furthermore, Poole et al.'s work assumes that the ontology provides the names of all properties and relations, but, much like MLN, restrictions on them are expressed probabilistically. Classes, on the other hand, seem to be rigidly defined. In contrast, our approach focuses on flexible definitions of classes while allowing purely logical restrictions if the data supports them.

6 Conclusions

We have outlined an approach to developing *Ontologies in Variation*: human-comprehensible representations of what is typical and variable among the instances of natural kinds, objects and concepts. Our formalism's novelty lies in the strategic complementation of axiomatic knowledge by statistical knowledge such that classical logical inferences are maintained while also precisely and explicitly describing variable aspects garnered from datasets.

Acknowledgements: We gratefully acknowledge intellectual contributions from Autodesk Research and funding from the Natural Sciences and Engineering Research Council of Canada. We thank the reviewers for their comments.

References

Bacchus, F.; Grove, A.; Halpern, J.; and Koller, D. 1996. From statistical knowledge bases to degrees of belief. *Artif. Intell.* 87(1–2):75–143.

Bacchus, F. 1990. LP – a logic for representing and reasoning with statistical knowledge. *Comput. Intell.* 6(4):209–231.

Demey, L.; Kooi, B.; and Sack, J. 2009. Logic and Probability. In Zalta, E., ed., *The Stanford Encycl. of Phil. (Spring 2013).*

Domingos, P.; Kok, S.; Poon, H.; Richardson, M.; and Singla, P. 2006. Unifying logical and statistical AI. In *AAAI-06.*

Federative Int. Programme on Anatomical Terminologies (FIPAT). 2011. *Terminologia Anatomica*. Thieme, 2nd ed.

Hahmann, T.; Samavi, R.; Mogk, J.; Bibliowicz, J.; and Khan, A. 2014. Towards generating bone-specific ontologies from an organic shape descriptor. In *Int. WS on Biomech. and Parametric Modeling of Human Anatomy (PMHA-2014).*

Hahmann, T. 2013. *A Reconciliation of Logical Representations of Space: from Multidimensional Mereotopology to Geometry*. Ph.D. Dissertation, Univ. of Toronto, Dept. of Comp. Sci.

Halpern, J. 1990. An analysis of first-order logics of probability. *Artif. Intell.* 46:311–350.

Hrdlicka, A. 1898. Study of the normal tibia. *American Anthropologist* 11(10):307–312.

Kuo, C.-L., and Poole, D. 2013. On integrating ontologies with relational probabilistic models. In *IJCAI-13*.

Mogk, J. et al. 2013. The parametric human project: building a probabilistic atlas of human anatomy. In *24th Congress of the Int. Society of Biomechanics*.

Poole, D.; Smyth, C.; and Sharma, R. 2008. Semantic science: Ontologies, data and probabilistic theories. In da Costa, P. et al., eds., *Uncertainty Reasoning for the Semantic Web I.* 26–40.

Predoiu, L., and Stuckenschmidt, H. 2010. Probabilistic models for the semantic web: A survey. In Tatnall, A., ed., *Web Technologies: Concepts, Methodologies, Tools, and Applications*. IGI Global.

Rosse, C., and Mejino Jr., J. L. 2003. A reference ontology for bioinformatics: the Foundational Model of Anatomy. *J. Biomed. Inform.* 36:478–500.

Schulz, S., and Hahn, U. 2007. Towards the ontological foundations of symbolic biological theories. *Artif. Intell. in Medicine* 39:237–250.