STANDARDIZING DATA DIMENSIONS FOR HEALTHCARE DATA WAREHOUSES

Richard E. Biehl, Ph.D.

University of Central Florida
Department of Health Management and Informatics

DIMENSIONS: AN ONTOLOGICAL PROBLEM

In a 1-dimensional warehouse, that dimension could list all of the necessary domain concepts; however, the low dimensionality would be insufficient to take advantage of the star schema dimensional model. A simple affinity principle could be used to split the one dimension into many dimensions: Divide up the concepts into groups such that the concepts within each share more in common with each other than with the concepts placed into the other groupings. Each cluster represents a potential perspective against which one might analyze or aggregate data in the warehouse: a *dimension*.

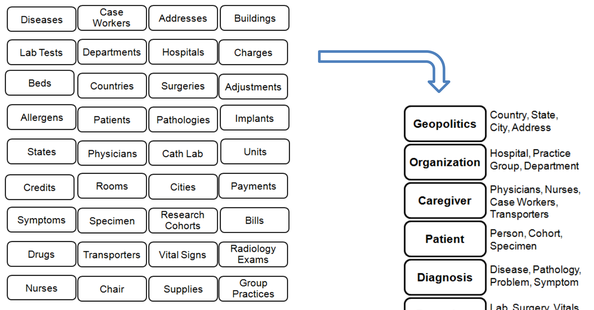

Fig. 1. Examples of logical groupings that might be identified as candidate warehouse dimensions during an affinity analysis of project-relevant data.

Fig. 2. Examples of physical dimensions with their associated logical subdimensions that might result from an affinity-based and heuristic analysis.

AFFINITY HEURISTICS

A number of useful heuristics has emerged for helping with the affinity analysis:

1. Clusters that represent systems of some form will typically reside in different physical dimensions if they represent systems of different scale: societal, organismic, or mechanical.
2. Clusters that differ in focusing on people, places, or things don't typically resolve into the same dimension.
3. Clusters that represent internal and controllable data don't typically map into the same dimension as external data that might be very noisy and out of our control.
4. Clusters representing physical things aren't usually mapped into the same dimensions as logical or conceptual things.
5. Clusters that represent things that one does (typically verbs) are usually mapped into a separate dimension from things that one has or possesses (typically nouns).
6. Clusters that represent attributes that describe something of interest typically end up in different dimensions than things that are used or needed for something of interest.

These heuristics are always subject to interpretation, and they won't always produce the same result when practiced by different warehouse development teams; but they go a long way toward reducing the high levels of variability in design that can be seen in dimension identification done using more *ad hoc* analysis.

DESIGN DILEMMA

Dimensional data warehousing involves the separation of facts and dimensions, with facts comprising the quantitative and observational data of interest, and dimensions comprising the variety of contexts in which those facts are collected and understood.

While a standard design pattern determines the structure of each of the various warehouse dimensions (Figure 8), it ultimately says nothing about the semantic meaning of those dimensions. The pattern is neutral with respect to defining a grouping of concepts into the dimensions.

The number and types of dimensions should be defined as naturally as possible from the problem domain and requirements being addressed by the implementation of the warehouse.

Fig. 7. Evolution of the "star" schema paradigm in data warehousing: (a) A traditional sales data view with a 5-point star (c. 1987), and (b) a much higher dimensional clinical data star with 15-20 dimensions (c. 2007

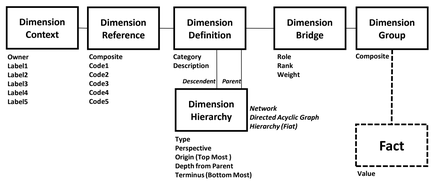

Fig. 8. Dimensional anatomy design pattern. *Bridge-Group* constructs allow individual facts to reference multiple entries in a single dimension, and *Context-Reference* constructs allow entries to be identified by the natural keys in multiple source or destination contexts. The *Hierarchy* construct allows for aggregation, disaggregation, and associations among entries independent of the local dimensionality of the facts.

BASIC FORMAL ONTOLOGY

The Basic Formal Ontology (BFO) is an upper-level ontology for the classification of real-world artifacts that serves as an effective general model for defining ontologically-oriented warehouse dimensions. At its highest level, the BFO divides objects into *continuants* and *occurrents*. All dimensions in a warehouse design should be traceable back to the BFO.

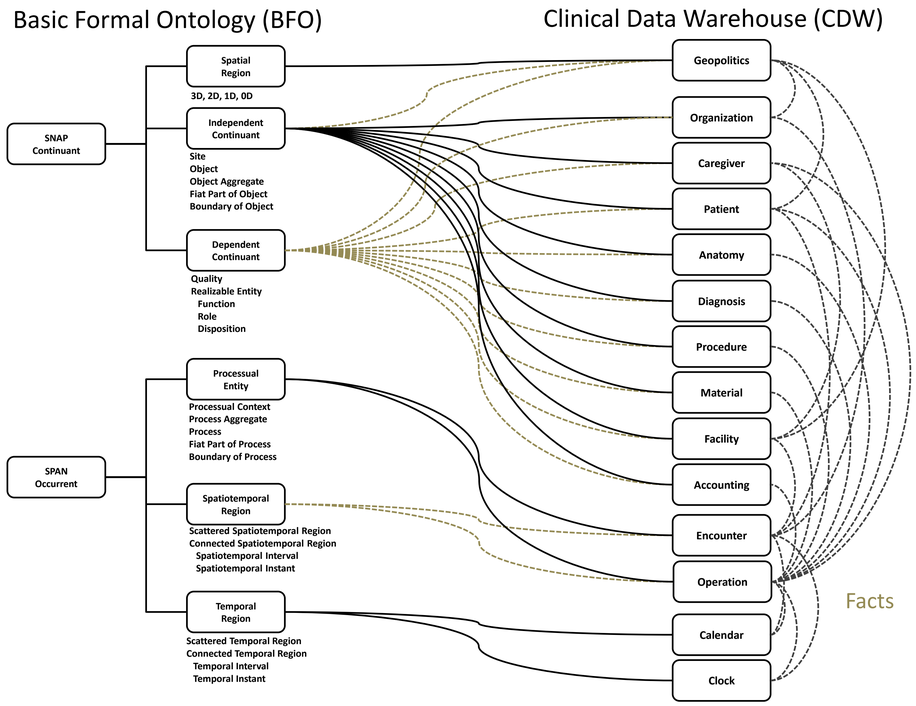

Fig. 3. Mapping of dimensions in the clinical domain against the classifications in the Basic Formal Ontology (BFO): (a) Calendar and Clock dimensions as BFO Temporal Regions, (b) Geopolitics dimension as BFO Spatial Region, and (c) all other clinical dimensions as BFO Independent Continuants.

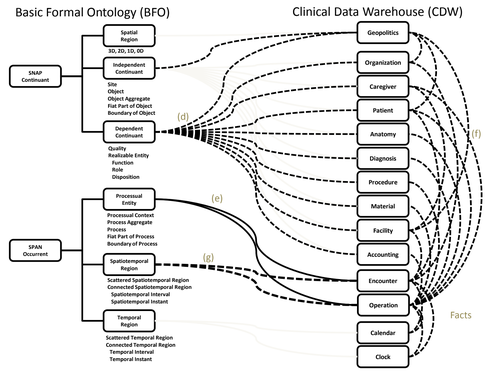

Fig. 4. Continued mapping of dimensions against the Basic Formal Ontology (BFO): (d) Properties of the clinical dimensions as BFO Dependent Continuants during Master Data Management (MDM) activities, (e) introduction of Encounter and Operation dimensions as BFO Processual Entities, (f) loading of facts using all dimensions for context, and (g) resulting interpretation of Encounters and Operations as BFO Spatiotemporal Regions (e.g., worldlines).

LOGICAL V. PHYSICAL DIMENSIONS

The dimensions of the data warehouse embody the set of concepts that provides the semantic context for all of the facts to be stored in the warehouse. Each distinct concept constitutes a logical dimension of the warehouse. Logical dimensions are typically collected into physical dimensions within the database technology that implements the warehouse. The number of physical dimensions is a *performance* concern, and the number of logical dimensions is a *domain* concern.

Early warehouse designs of low dimensionality can be changed to become higher-dimensional warehouses without losing the domain perspective supported by the logical dimensions contained in those physical dimensions. Each warehouse will vary along a continuum (Figure 9) from one extremely generalized dimension that can define and store any dimensional reference, to a collection of very specific dimensions that each stores definitions that share semantic context and meanings within the problem domain.

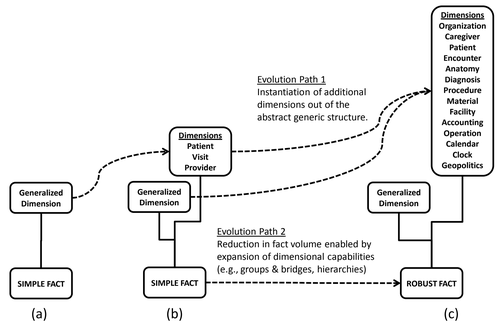

Fig. 9. Dimension Maturity Pathway: (a) 1-dimension warehouse with all logical dimensions generalized into the single physical dimension, (b) the i2b2 design, with three specific logical dimensions moved into their own physical dimensions, and (c) a clinical warehouse with 14 physical dimensions, and everything else left in the generalized dimension. The generalized dimension is always present, and embodies any logical dimensions that don't map into one of the defined physical dimensions. The number of logical dimensions remains constant.

INCORPORATING OTHER ONTOLOGIES

Once the BFO has been used to model warehouse dimensions, a more practical question becomes: How can other ontologies better inform the definition of our dimensions? Mapping publically available ontologies (Figure 5) into the warehouse design provides for analytical capability in the warehouse that was unknown or unavailable to the warehouse's source systems. The dimension design pattern provides standard constructs for representing the relationships defined within these ontologies without any database modifications.

Ontology	Ontology Element	BFO Element	Dimension
Foundational Model	Dimensional Entity	Spatial Region	Anatomy
of Anatomy	Anatomical Entity	Independent Continuant	Anatomy
	Attribute Entity	Dependent Continuant	Anatomy
Human Disease	Disease	Independent Continuant	Diagnosis
Human Phenotype	Inheritance	Independent Continuant	Diagnosis
	Organ Abnormality	Independent Continuant	Diagnosis
	Onset & Clinical Course	Dependent Continuant	Diagnosis
Environmental	Environmental Matter	Independent Continuant	Material
	Food	Independent Continuant	Material
	Biome	Independent Continuant	Facility
	Environmental Feature	Independent Continuant	Facility
Gene	Cellular Component	Independent Continuant	Anatomy
	Biological Process	Independent Continuant	Procedure
	Molecular Function	Independent Continuant	Procedure

Fig. 5. Mapping of public ontologies against BFO-based warehouse dimensions.

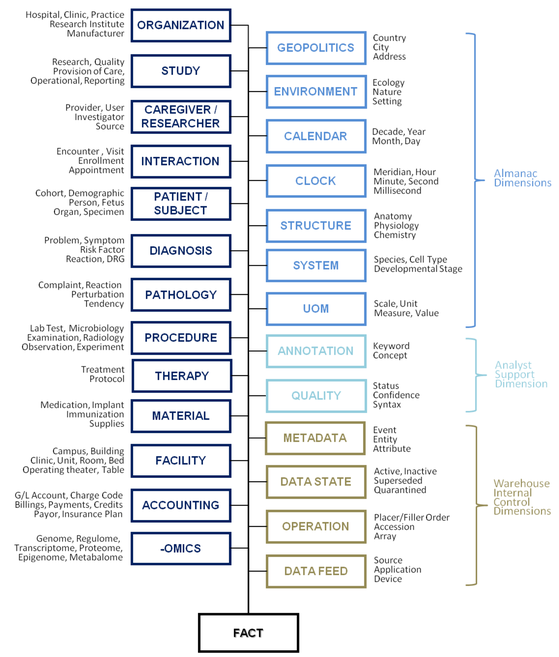

Fig. 6. An example healthcare data warehouse built using BFO-validated dimensions

LIMITATIONS

Following these heuristics can produce a range of possible design alternatives, with the less appropriate or incorrect designs largely excluded. But the choices that are required by the design analyst in selecting from among the better alternatives still involves numerous judgment calls that require a good knowledge of the warehouse requirements and design experience in making such selections.

CONCLUSION

Ontology-based dimensional data warehouse design is a set of heuristics that can greatly improve the quality and effectiveness of data warehouses. While meeting organizational requirements and database design principles needs to be paramount, aligning the dimensions of a data warehouse with good ontology practices can greatly enable a warehouse to meet more extensive requirements for the future, and greatly increase semantic interoperability with other data warehouses in other organizations that are working to align with the same ontologies. Actual data in the warehouse, when analyzed against known ontological alignments, will improve confidence in existing ontology relationships, and suggest new relationships not yet recognized.