

Slowly Changing Dimensions: a Pattern Language for Coping With Change in Analytical Information Processing

Hans Wegener*, Robert Marti†

June 5, 2007

Abstract

In many larger organizations like internationally operating companies, owners of analytical information systems must frequently cope with change. An important part of this problem is the restructuring of attributes serving as classification categories, so-called *dimensions*. This pattern language concerns itself with different ways of preparing for and dealing with changes in the real world and their impact on the analysis of *derived measures* based on historic data in *time-series*.

Note to EuroPLoP 2007 Sheperds and Reviewers

This pattern language was started at EuroPLoP 2005, when it contained the patterns in Section 4. While the entire language is still subject to revision, we are currently most interested in receiving feedback on Section 3.

1 Motivation

In any enterprise, conducting and closing business transactions such as selling products and services, buying supplies, etc. leads to the collection of business facts that measure things such as volume, speed, and quality of the business conducted. These business facts are in turn used to calculate various performance indicators (e.g., profit, margins, return on capital) and risk indicators (e.g., volatility of investments) which serve as one of several inputs to steer the future direction of the enterprise, including establishing targets and limits for the next business cycle.

The connection between collected “raw” measures in business facts and performance indicators can be quite intricate, involving complex calculation steps. In a large enterprise with partly autonomous business entities, this complexity is compounded when business facts collected in various business processes and/or locations only match partially, e.g., due to different encodings of

*Zurich Financial Services, P.O. Box, 8085 Zürich, Switzerland, Hans.Wegener@zurich.com

†Swiss Re, Mythenquai 50/60, 8022 Zürich, Switzerland, Robert_Marti@swissre.com.

currencies, countries, business lines etc. In addition, the business environment is changing over time, e.g., due to the development of new products, changes of customer characteristics, and internal reorganizations, making it difficult to reconcile and compare business facts collected over the years.

In the following, we outline an approach to cope with these issues, especially dealing with change in the business environment over time. We shall focus on changes to attributes which categorize business transactions, which, following [5, 6, 9], are called dimensional attributes, given that these attributes define an n-dimensional space or hypercube for the collected measures (see also below)¹. Such changes demand specific treatment because they often affect time-series² analysis of performance indicators. Imagine that a business application deals with an abstraction of Germany. When the notion of Germany changes (as happened with reunification in 1990), some of the existing data managed by the application (such as old gross national product figures) may have to be adapted to the new notion. The same is true for former Yugoslavia, Czechoslovakia, Congo, etc., as well as for changes in the profit center structure of a company and changes to the structure of companies caused by mergers and acquisitions.

The execution and closing of business transactions—selling a product or service, buying supplies, signing and terminating employee contracts etc.—is the main source of business facts that need to be recorded in an enterprise. Further data sources external to the enterprise—mostly changes in the geo-political and economic environment such as the rise and fall of inflation, interest rates, mergers and acquisitions of clients and competitors, etc.—need to be tracked and recorded as well. A fact can be viewed as a record that typically consists of a series of attribute values including

- A unique identifier to distinguish this business transaction from other transactions.
- A transaction date stating when the data pertaining to the transaction was recorded.
- An effective date stating when the result of closing the transaction is effective, e.g. the time period during which an insurance coverage is in effect.
- Measures which indicate quantities or monetary amounts involved, e.g. total sum insured, deductible, premium due, base salary, etc.
- Dimensional attribute values that describe the context of a business transaction, e.g. the identification of the business partner, the product or service offered or sold, the responsible profit center, etc.

For some measures, we do not know (or care) how their value was arrived at—whether they were computed from other data, are the result of a measuring or counting process, represent an estimate, or were obtained from a third party. Such a measure is called an observed measure. In

¹Given that changes to dimensional attributes (addition and removal as well as hierarchical restructuring of the values of such an attribute) is relatively slow with respect to the rate at which business transactions are captured, the issues related to the change of dimensional attributes has come to be called “slowly changing dimensions”.

²A time-series is a sequence of data points, measured typically at successive times, spaced apart at uniform time intervals. Time-series analysis comprises methods that attempt to understand such time-series, often either to understand the underlying theory of the data points, or to make forecasts [15].

contrast, a derived measure is computed from observed measures and other derived measures, e.g., by applying a formula describing how it is computed.

For example, at least for the purpose of information integration and analysis, the annual base salary of an employee is usually treated as an observed measure—even though the annual base salary might actually have been computed according to a complex formula which takes other measures and parameters into account, e.g., highest educational degree reached, age, number of years experience in the field, rank, and number of years employed. On the other hand, total annual compensation, computed from the observed measures annual base salary, family and child allowance, and annual bonus, is a derived measure.

Dimensional attributes³ are attributes which describe the context of business transactions: They refer to objects such as business partners, products or services, (internal) profit centers, responsible employees etc.⁴

Dimensional attributes typically assume values taken from a pre-defined set of values or, somewhat more dynamically, a set of object identifiers for specific objects. In the following, the structure of a dimensional attribute is considered to consist of the set of values that belong to the dimensional attribute.

2 Pattern Language Overview

We distinguish five blocks of activities in the context of this pattern language, namely

- *Deciding how to structure dimensions*: addresses the issue how, given the multitude of modeling options and statistical constraints, categories should be structured to achieve stated analysis goals. This may be considered the static perspective.
- *Deciding how to look at changes*: concerns itself with the question how, given a change in the real world, it should be reflected in the business model on which the IT systems are based. This may be considered the dynamic perspective.
- *Separating responsibilities in the application architecture*: decides how the tasks involved in handling a change are divided on the technical or architectural level to maximize reuse of functionality. This may be considered the strategic perspective.
- *Propagating a change through the application architecture*: plans, given a separation of responsibilities in the application architecture, the steps necessary to propagate it throughout the corporation, minimizing the incurred operational risk. Executes the plan. This may be considered the tactical perspective.
- *Deciding how to historize data*: relates to the implementation of historization requirements in technical systems given formulated constraints on cost, performance, and auditability. This may be considered the technical perspective.

³Instead of *dimensional attribute*, papers on statistical analysis often use the term *categorical attribute*.

⁴In rare cases, an attribute may play different roles in different transactions: for example, when underwriting an insurance contract, the contract identifier helps to identify this business transaction, together with the year of coverage. This same contract identifier will be used as a dimensional attribute for claims processing, as every claim must refer to an existing insurance contract.

The main relationships between the above activities are depicted in Figure 1. Earlier work already addressed some of these issues, mostly from the implementation perspective. Arnoldi et al. [2] discuss various techniques to store an object's history and present it to users from different perspectives. Fowler [5] goes a little further in explaining the details of storing history, as does Anderson [1]. All three allude to the issue of bitemporal storage, which is discussed in depth by Rüping [10] and, above all, Snodgrass [11].

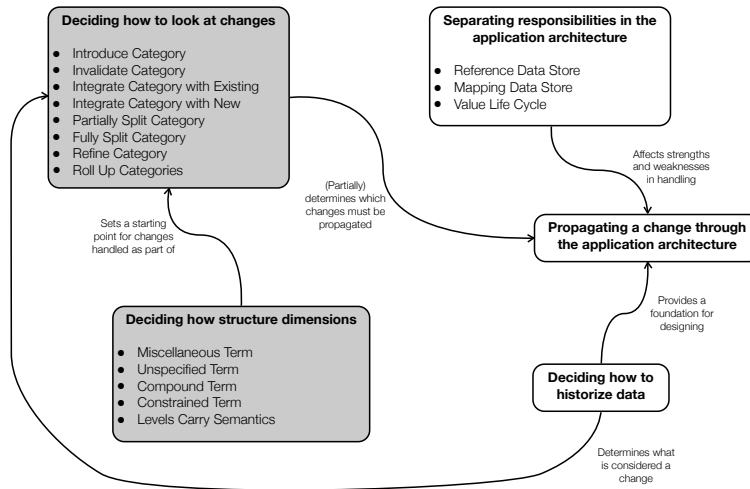


Figure 1: The shaded area outlines the scope of the rest of this paper.

3 Deciding How to Structure Dimensions

Business terms represent concepts in the real world. Dimensions are used to group such terms into more generic and more specific categories. These categories are subsequently used to classify business transactions according to their characteristics. Given enough time, administrative staff can identify the business term most appropriate for describing a transaction by reading its definition and comparing it to others. In today's typical cost-conscious corporate environment, however, such a luxury is not afforded many an employee. Instead, limited time is available to assign a category to a transaction. Besides, providing concise definitions is a costly business as well, taking up hours of expert work time that is often not adequately rewarded. In the absence of organizational sources of knowledge like expert advice, the name and structural relationships of a business term are often the only other way to help administrative staff to choose categories wisely. Hence, the problem arises

Which conventions can be used to communicate the business meaning of a category, complementary to its definition?

On the other hand, statistical analysis of (historic) data requires it to be not just correct but also comprehensive. Business controllers and other specialists who sift through large amounts of transactions depend on administrators to classify them in a way that is not only appropriate, but exact. Some of the shortcuts taken by using conventions on naming and structural relationships can

dilute that exactness. The classic tradeoff in this context is between the administrator’s priorities (get the job done quickly) and the analyzer’s needs (get the job done precisely). When designing a dimension with its values, this tradeoff must be made, and the naming and structural relationships of business terms is faced with yet another problem, namely

How do you structure a dimension such that the diverging needs of administrative and analytical staff are in balance?

This pattern language concerns itself with solutions to these two problems. It uses concepts in the spirit of EVOCATIVE PATTERN NAME [8] in that it tries to use means other than a concise definition to convey the meaning of a business term. The thumbnails in Table 1 give an overview of the patterns to be presented.

Pattern	Problem	Solution
Miscellaneous Term	How do you represent the complexity of the real world in the children of a dimensional value?	Limit the number of children to about seven, each of which with a crisply defined scope easily recognized within the context of the parent term. Add a term named "Miscellaneous" or "Other" that covers the remainder scope of the parent.
Unspecified Term	How can you indicate that a classification has not (yet) been made?	Provide a term named "Unspecified," "Undisclosed," "Not Applicable," or "Can't say" at the topmost layer of your dimension.
Compound Term	How do you name a parent term that groups (seemingly) unrelated terms?	Concatenate the names of the child terms and separate them by the symbol "&" and use that as the parent term's name.
Constrained Term	How do you eliminate a degenerate branch in a dimension?	Collapse the terms in the branch into one by concatenating their names, separated by the symbol "/", replacing the (topmost) parent value.
Levels Carry Semantics	How can you reify more than the generic-specific relation between values in a dimension?	Require values with a common property be put on a specific layer in the dimensional tree, and associate that property with the layer.

Table 1: Thumbnails for patterns helping you to decide how to structure categories.

3.1 Miscellaneous Term

In a dimension, values grouped below another value are considered more specific terms than that of the parent term. As such, when transactions are recorded and classified (“booked”), these children offer a selection, a one out of many choice. The task of the administrator is to pick the most appropriate from this selection. However, without prior knowledge about the domain at hand it can become tricky to choose the terms that should become part of the offering. In addition, administrators can become confused when there are too many values to select from.

At the other end, people concerned with analyzing business results in the steering process require the classification system to be as precise as possible. Ideally, each term would have an exactly specified meaning, and it would be easy to identify the one and only way to classify a given transaction. As such, the selection should be as fine-grained as possible, leading to many business terms subsumed under the parent value and a very “deep” dimension with many layers of terms.

The obvious challenge is to find a sweet spot between the needs of administrative and analytical staff, that is the right number of children for a given parent value. What is more, these child terms must be named in a way that administrators can pick them easily, while it must be ensured that booked data can be used effectively to discover statistically relevant effects. Hence, the question arises

How do you represent the complexity of the real world in the children of a dimensional value?

Adding a child term for every conceivable category in the real world may permit administrators to identify the “right” choice (assuming there is such a thing at all). However, that would make the task of selecting it too onerous, given the many children there might be.

Humans are good at recognizing and filtering up to seven, maybe nine concepts at once. Anything beyond that requires training or experience. Parent terms with more than ten children will likely overwhelm administrators, leading to bookings that are in error. However, there are cases where permitting more than ten children is indeed necessary.

The lower the number of children, the more difficult it is to discover statistical aberrations, because there is no classification that sets them apart from others. In extreme cases they can go undetected. However, too many children may cause groupings that are not meaningful, as they reduce the number of bookings, and thus their statistical relevance.

Identification of categories is a lengthy process with occasional reversal of previous decisions. Taking design decisions on a categorization scheme too early may lead to costly repair work due to re-categorization and the consequential re-computation of derived measures. Hence, you will on the one side be inclined to have a stable set of categories, yet allow for some flexibility to change your mind.

Therefore:

Limit the number of children to about seven, each of which with a crisply defined scope easily recognized within the context of the parent term. Add a term named “Miscellaneous” or “Other” that covers the remainder scope of the parent.

Using MISCELLANEOUS TERM means that administrators have an easier time deciding if they know exactly what to choose (suggesting a selection of one of the crisply defined terms) or whether they know that they don’t know (suggesting choice of the generic term).

With this it becomes possible to wait with adding new categories as children, as they can safely be classified under MISCELLANEOUS TERM. As more stable categories are identified, less and less require to be classified there, but given a crisply defined and well-understood home.

Using MISCELLANEOUS TERM in the presence of too few peer terms (usually three and less) waters down the ability to detect statistically relevant events, as they are now “hidden,” categorized among many others in MISCELLANEOUS TERM.

Because it can be used at each level in a classification tree, prolific use of MISCELLANEOUS TERM can result in term bloat. Forcing it to be offered below all non-leaf values in a dimension may inundate its designers with tedious work. It may also prove superfluous.

Terms like “Miscellaneous” or “Other” as used in MISCELLANEOUS TERM must be distinguished from the term “General,” which has a slightly different meaning. The latter often refers to a category that covers the scope of all other terms (below the same parent), as in a foundation. The JEL [7] category C0 (Mathematical and Quantitative Methods: General) or the ACM classification A (General Literature) are such cases.

In one of the companies the authors have seen, the Line of Business contains, among at least two dozen such cases, the terms Annual Risks Miscellaneous (under Engineering Annual Risks), Liability Miscellaneous (tellingly grouped below Other Liability), and Non-Life Disability Miscellaneous (as part of Non-Life Disability).

The WHO International Statistical Classification of Diseases [17] features a category I95-I99 (Other and Unspecified Disorders of the Circulatory System) below the parent term I00-I99 (Diseases of the Circulatory System). Interestingly, below the term J00-J99 (Diseases of the Respiratory System) you find two cases of MISCELLANEOUS TERM, namely J90-J94 (Other Diseases of Pleura) and J95-J99 (Other Diseases of the Respiratory System).

The ACM computing classification B.m (Hardware Miscellaneous) is such a case in point. In fact, all top-level categories have each two child terms, one Miscellaneous and one General.

3.2 Unspecified Term

Sometimes data is only available incompletely. An administrator may be faced with a form that has just been filled out half-way, but needs to be captured right now to bring another process forward. Think of someone arriving at a hospital by ambulance: it is inconceivable that the patient would only be rushed to the operation room after the person's meal preferences (vegetarian, kosher, or regular) are captured.

Another phenomenon in statistics is that of forced choice: offered three different categories to choose from, an administrator (or, for that matter, a respondent to a questionnaire) will always try to pick the one that seems most reasonable—even if that means bending the truth. Forced choice can distort data distribution by introducing an element of randomness in the activity of classification.

In some industries such as financial services, privacy concerns play a very big role. They can counteract the need for fine-grained statistical classification. In health insurance, data is sometimes anonymized or obfuscated, requiring a special concept to represent this state. Hence, the question arises

How can you indicate that a classification has not (yet) been made?

Especially if the dimensional hierarchy is deep, an administrator can select a MISCELLANEOUS TERM close to the leaves of the tree, if available. That term would then adequately, although not precisely categorize the transaction. However, a person analyzing the data may not be aware of this fact and take the classification as accurate, which it in fact is not.

Database designers sometimes make use of the NULL value to mark entries not available. However, the intricacies of relational query logic then force them into all sorts of special treatment for this obscure value, which is handled differently in boolean operators, (inner) joins, and more.

If the administrators just mark a booking as “not classified,” that also means that it may not be considered by some analyses. This can, in extreme cases, lead to the exclusion of a significant portion of the data set, rendering the analysis statistically irrelevant.

Therefore:

Provide a term named “Unspecified,” “Undisclosed,” “Not Applicable,” or “Can’t say” at the topmost layer of your dimension.

The use of this pattern helps analytical staff to distinguish booked facts that can be used in statistical analysis from those which should be excluded (from the perspective of a specific dimension, that is), while administrators can avoid a forced choice.

Prolific use of UNSPECIFIED TERM may render the data set unusable for particular statistical analyses. Care must be taken to educate administrative staff in this regard, for example in the form of an awareness campaign.

You will be forced to refrain to other measures (such as MISCELLANEOUS TERM) in cases where you know something about the characteristics of the transaction, but not all.

In the WHO International Statistical Classification of Diseases [17], under Diseases of the Musculoskeletal System and Connective Tissue, an optional subclassification is provided to indicate the site of involvement. One of them is coded 9 (Site Unspecified).

In one bank the authors know, the application landscape is managed for resistance to change by tracking the alignment of applications with the different platforms offered for the various architectural layers. Less alignment is assumed to lead to a higher resistance to change, so the type of alignment of each application in the portfolio matters. It was found that besides the regular classification of alignment (Full, Partial, None) required an additional category (Not Applicable) to cater to cases where the rules of the architectural specification did not make much sense.

3.3 Compound Term

It can happen that a term acts merely as a grouping mechanism for a bunch of loosely or unrelated terms. This might be because taken together they fulfill a particular (statistical) purpose; it might also be that the child terms are left over and just need a place to stay. The more of such children there are, the more difficult it will be to find a common term to convey the meaning of the group appropriately.

Concatenating the child terms to form the name of the parent term works only for two, maybe three terms. Beyond three children the parent name becomes very difficult to parse and other means should be sought. In this situation, the question arises:

How do you name a parent term that groups (seemingly) unrelated terms?

You can just give it a short name of your own choosing. However, that often fails to guide the thoughts of administrators: the more generic concept represented by the parent term may, fail to capture the attention of people using them or be misunderstood. This may result in erroneous classification during capture or interpretation during analysis.

With a decreasing common semantic denominator among the group of child values, finding a term name that describes all of them becomes increasingly difficult. Just enumerating the child values' names, can be a way out. However, as their number grows, it becomes just as confusing to the human reader.

Therefore:

Concatenate the names of the child terms and separate them by the symbol “&” and use that as the parent term’s name.

If the group of terms is large, but has little or no common denominator at all, you can choose MISCELLANEOUS TERM as parent value for an alternative way out.

Human language does often not follow the rules of boolean logic. The symbols “&” or “/” can be used to signify the logical operator “or” despite the fact that people speak of, say, “Property & Casualty” when really meaning the inclusive “or.”

People sometimes use the symbol “/” in different senses—once signifying “or,” other times meaning “and.” There are also situations where it is used to separate the specific from the generic

part of a term name, as in “Toys/Children/Wood”.

The concatenated term will be more sensitive to changes in the semantics, names, and composition of its child terms, as for example in case of INTEGRATE CATEGORY WITH EXISTING, INTEGRATE CATEGORY INTO NEW, PARTIALLY SPLIT CATEGORY, and FULLY SPLIT CATEGORY.

Three examples from insurance are the business lines Property & Casualty, Life & Health, and Aviation & Space. They group the traditional non-life and life lines, respectively, as well as the business of Aviation insurance (air planes) and Space insurance (satellites, rockets).

3.4 Constrained Term

Classification systems try to group business terms into more generic and more specific categories to allow for drill-ups (aggregation of computed measures across a group of categories) and drill-downs (selection of a subset of facts categorized in one particular way to analyze their computed measures).

The idea is that each parent term has at least two child terms. It would not make sense to add but one child term, since drill-down and drill-up would essentially yield the same computed measures (assuming that only leaf-level categories are booked). However, there are situations where it indeed happens that but one child category ends up grouped below a parent. For example, the company employs an external standard like ISO, SWIFT, FIX, or ACORD but chooses to use only parts of it; or the data is found only to require a single category.

Consequently, the dimensional tree may sport a branch that is actually a chain of values, each having exactly one parent or child value, respectively. These shall be called a degenerate branch. This is superfluous, for which reason we are permitted to ask:

How do you eliminate a degenerate branch in a dimension?

In order to keep business terms easy to understand, you would prefer to see each be composed of a single noun, perhaps qualified by an adjective or preposition, such as Professional Liability. However, that can be difficult to achieve when parent and child(ren) are all nouns or activities, such as “Trading” (parent) and “Securities,” “Bonds,” “Derivatives” (children).

Therefore:

Collapse the terms in the branch into one by concatenating their names, separated by the symbol “/,” replacing the (topmost) parent value.

Typically you choose to put the most generic (former child) term first, and the most specific one last. Humans expect this ordering, which makes it easier to parse.

In natural language, the symbol “/” is used for other purposes as well, such as in composition (see COMPOUND TERM). Furthermore, comma and whitespace are sometimes used as mechanism to constrain a term. In the latter case, however, very often the parent name is concatenated with the specific term, such as in Allocated Investment Return for Reinsurance, Retrocession.

The resulting term relieves you from an artificially deep nesting of terms. However, it can become difficult to parse, especially when they are composed of more than three names that do not follow the rules governing the natural language they are designed for.

In one data language we have seen, the following terms can be spotted: Stop Loss/Aggregate XS (non-proportional cover offering aggregate protection in a specified class of business for the total annual loss burden in excess of the insurers deductible up to the agreed limit) or Marine

Fine Arts/Exhibitions (transportation and exhibition as well of art, antiques, or samples by marine vessel).

3.5 Layers Carry Semantics

A dimension offers categories to classify transactions from a particular point of view. The generic-specific-relationship between parent and children makes the taxonomy from that point of view explicit, allowing you to computed (aggregated) measures. However, it does not guarantee (yet suggest) that any other form of organization takes place. There is no guarantee that child terms share any other semantic property beyond being more specific than their parent. Consequently, if you must classify transactions from two or more perspective, you are at a loss.

Requiring more than one attribute to classify transactions from more than one point of view may not be feasible, however. You might be using an off-the-shelf software product that allows for only one form classification, or you might be struggling with legacy systems that share the same limitation.

How can you reify more than the generic-specific relation between values in a dimension?

Often, classifications are mapped to some form of organization, which again often follows regular patterns. For example, the CEO is head of the executive board, the members of which are boss of the executive directors, and so on. This regularity can be exploited. However, regularity that is not managed as a reified concept may suddenly cease to be. Sudden changes may bring a mechanism exploiting it to an unwelcome halt. Database queries may stop working or analyses yield strange results.

Therefore:

Require values with a common property be put on a specific layer in the dimensional tree, and associate that property with the layer.

Organizing business terms like this makes the dimension more sensitive to changes. For example, one company we know used to be organized by geographical region (Europe, Asia, Americas), then it changed to a product-oriented organization (life and non-life insurance plus financial services). Recently it changed to a process-oriented setup (client markets, product development). Each time the semantics of layer 1 of the line management structure changed, requiring subsequent changes in other places.

Sometimes administrators work only with the values at a certain level, such as the controller at the corporate center who is only responsible for the company's business units, but not its regional operations. In that case it can be observed that the values from a particular layer are put into a separate dimensional attribute, so that administration is more convenient and prevents bookings at a "wrong" level.

Assigning meaning to layers makes it easier for administrative and analytical staff alike to book and analyze data, because there is now additional, and partially redundant indication what the categories mean in terms of the business.

Most of the time the common property will be simple, such as "is an animal" or "requires additional checks." Nothing holds you back, however, to come up with a scheme where values at layer $0 - n$ have the same property, while values at layers $m > n$ each have other properties.

Using this pattern can lead to an increasing number of homonyms. For example, the profit center structure in one of the companies we know contains the layers Division, Market, and so on.

Certain contracts or business associated with special purpose entities is booked on profit centers at, say, layer four. Since they are so special, there is no natural parent profit center candidate, for which reason a number of artificial, like-named categories are created that link the special profit center to the topmost category. Such naming schemes can become confusing, for which reason booking is often restricted to the leaves of the dimensional tree (to make category selection easier).

Using this pattern can simplify database queries. For example, to find the regional outliers (leaders, laggards) you may only have to select data associated with categories on a certain layer—given that all regions are put there. In fact, many, primarily older database applications use coding schemes for business terms in dimensions, such as (regular expressions used):

- Level 1: $code = [0 - 9]^*$
- Level 2: $code = [0 - 9]^*[0 - 9]^*$
- etc.

The business terms at level 2, for example, can then be extracted by querying for the terms with codes of length 2. As can be easily seen, however, such coding schemes restrict the number of terms that can be placed below a parent (10 in this case).

The JEL classification system [7] requires classifications be made using the categories at level two of the classification hierarchy. This leads to homonyms—even in different branches of the classification tree, such as Schools of Economic Thought and Methodology → General → General, Mathematical and Quantitative Methods → General → General, and others.

The Linnaean taxonomy [14] arranges species in the following hierarchical order: Superregnum → Regnum → Subregnum → Superphylum → Phylum → Subphylum → Infraphylum → Superclassis → Classis → Subclassis → Infraclassis → Ordo → Subordo → Familia → Subfamilia → Tribus. For example, the (house) cat is classified as Eukarya → Animalia → Eumetazoa → Deuterostomia → Chordata → Vertebrata → Gnathostomata → Tetrapoda → Mammalia → Theria → Placentalia → Carnivora → Feliformia → Felidae → Felinae → Felis → Felis sylvestris → Felis sylvestris catus.

4 Deciding How to Look at Changes

Events within and outside an enterprise may lead to the decision to change how it looks at the world. As an effect, the structure of a dimensional attribute will (or will not) be changed. For example, due to administrative changes in the 1990s, Hong Kong and Macao became provinces of the People's Republic of China and ceased to be nation countries. However, from an actuarial perspective (e.g., making reserves for expected insurance claims), the two continue to be treated differently from mainland China. The question must be answered as to if, and how to reflect the real-world change in the virtual-world model, which in this case happens to be the dimensional attribute Country.

Such changes carry the potential to affect (derived) measures computed at an earlier point, i.e. based on a historic attribute structure. Which structural change to an attribute is taken as commensurate to the business change in the world thus determines how the amended view must be computed. One common problem therefore is to understand

What is the impact of a structural change to a dimensional attribute on the amended view, i.e. how do (derived) measures change, if at all?

Changes can become rather complex; comprehensive, composite changes may be just as appropriate to the change as a group of simple, atomic ones. However, the way composite changes are put together may affect the way history is looked at. Once (at least) two alternatives have been identified, it becomes beneficial to know

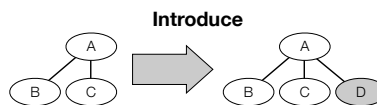
What are the consequences of different courses of action, i.e. different ways of looking at a given change? How does this affect the analysis of historic data?

We assume for all changes that knowledge of them is obtained before their date of effectiveness. This simplification holds true for most of real life. Furthermore, should it happen that one learns of a change after the fact, adopting it follows the same line of thought and course of action, only the impact may be different. The pattern thumbnails are shown in Table 2.

Pattern	Problem	Solution
Introduce Category	Which way do you introduce (genuinely) new categories in the face of previously categorized business data?	Add the new category to the dimensional attribute and give it the life cycle status active. Ignore historic data that might be categorized the same way.
Invalidate Category	How should you treat values that have become obsolete?	Change the life cycle status of the value to inactive.
Integrate Category with Existing	Which survivor do you choose to replace outdated values, how do you time your change and how do you map outdated values?	Choose a value that has a close resemblance to the outdated value(s) and remove them from the attribute; change their life cycle status to inactive or superseded; map the old values to the survivor.
Integrate Category into New	What is a good choice of successor to replace the outdated values and how do you map them?	Identify a business term encompassing all outdated terms. Remove the value(s) to be phased out from the attribute and change their life cycle status to inactive or superseded; map old values to the survivor.
Partially Split Category	How do you choose the new values and how do you map existing data?	Add the new values to the attribute in the reference data store, life cycle status active, map the old value to the new values.
Fully Split Category	How do you choose the new values and how do you map existing data?	Add the new values to the attribute, life cycle status active, and map the old value to the new values. Remove the old value from the attribute and mark it as superseded.
Refine Category	What do you do when you need to understand more detail about bookings of a certain class?	Add a layer of children below the parent currently classifying the bookings. Re-categorize them ("refine") as appropriate using those child terms.
Roll Up Categories	What do you do when a parent term has less than three children?	Eliminate the children and re-categorize ("roll up") all bookings with the parent term.

Table 2: Thumbnails for patterns helping you to decide how to look at changes.

4.1 Introduce Category



Choosing the most useful categories in order to gain valuable insights when analyzing business data is a constant challenge. Categories may need to be introduced either because they are completely new in nature, or because their importance has increased so that they should not be subsumed by another existing category any longer. You may recognize patterns emerging in your data and want to understand them better: for example, you might want to know how profitable a particular product is or what customer segment is most interested in it. In order to perform such an analysis, you must introduce a new category and start recording data based on it.

Which way do you introduce (genuinely) new categories in the face of previously categorized business data?

It may take you some time to decide to introduce a new category you may not recognize the new pattern early enough, or for lack of its importance you may have decided to tentatively ignore it. Therefore, by the time of making the decision, data will have been recorded that, in the light of the new category, would have to be re-categorized. Since the information required to perform such a re-categorization may have to be obtained manually, or worse, be irrevocably lost, this can become rather costly. Depending on how late you introduce a new category, transactions conducted previously may not be appropriately categorized. They will either have to be re-categorized, or reported measures will not give a fully accurate picture of history.

Re-categorization is a painful, expensive, and not completely unambiguous process. You will want to avoid it. However, the cost must be weighed against the benefits, which are sometimes difficult to assess.

A new category can be part of a bigger (i.e., composite) change that entails more than just one such category. Take the case of reorganizations: some of the organizations being introduced have been designed to replace others. These cases must be treated differently to ensure historic continuity of measures computed for previously existing organizations.

Therefore:

Add the new category to the dimensional attribute and give it the life cycle status active. Ignore historic data that might be categorized the same way.

Data recorded before introduction of the new category is not re-categorized, which may be inappropriate from the latter's perspective. But the change can be released very effectively, because it has no impact on measures recorded previously. No history needs be restated, no measures recomputed.

Introducing a category Nanotechnology, Computer Hacking or Genetically Modified Organism to the attribute Type of Peril in commercial liability insurance is an instance of this pattern: it would be difficult to claim that many existing transactions (i.e., facts) were associated with other values which would now have to be transferred to the new ones. Hence, once these perils are covered in a company's insurance contracts they can be added to the attribute describing them.

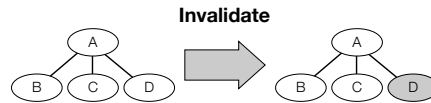
The German information service Heise Online reported in March 2005 [4] that Lycos had observed a 400% increase in the volume of unsolicited e-mail offering Luxury Goods over the past year. This new Type of Spam would now have to be added to the likes of Get Rich Quick, Phishing, Health Pills, or Sexually Explicit. Interestingly, as of March 2004 anti-spam tool vendor Sophos [12] only recognized the spam classes Offensive, Scam, Entertainment, Financial, Health, Internet, and Other content.

The ACM Computing Classification System [3] introduced a new category High-Speed Arithmetic in 1998 that did not exist beforehand. Within the category Arithmetic and Logic Structures

there were four categories before 1998, none of which had any resemblance to the former.

Typically you introduce a new category on a green field. However, it may happen that it has to be re-introduced. In this case it must be ensured that it has no (longer) business facts associated.

4.2 Invalidate Category



Things go out of fashion, cease to exist or slowly lose importance. At one point or another you may find that they have lost their significance for your analytical purposes. Two questions will concern you: how to deal with the outdated business category itself in terms of its use for time-series analysis, and how and when to remove it from its attribute.

When a value is replaced by (one or more) others, historic data associated with this value will be looked at in a different way. However, sometimes successor values are not easily identified. Forcing restatement of such data (i.e., choosing successors at all cost) may lead to reporting of distorted figures, as the underlying facts do not really belong where they are claimed to belong. This would be confusing, for which reason you may decide to discontinue the time-series analysis for this particular value. Once you have decided so, you must prevent the creation of new business facts associated with that value, which would be confusing as well.

How should you treat values that have become obsolete?

Clutter should be gotten rid of. However, business people often demand the continued use of outdated values for a long time. It can be difficult to get rid of such values.

The larger the volume of business recorded for the outdated value, the more likely it is that you must find a successor so you can account for it in your analysis. Finally, some data stores may continue to record transactions based on it.

Therefore:

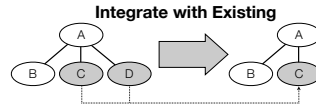
Change the life cycle status of the value to inactive.

You still have to actively manage the elimination of the invalidated category (i.e., complete removal from the attribute). Based on the life cycle status, data stores can prevent new facts for the outdated value to be added. Historic data associated with the value will ultimately be lost for time-series analysis. No history needs be restated, no measures recomputed.

When VW introduced its Golf car in 1974, the Beetle was still being produced. While some hailed the Golf as the successor to the Beetle, it was a strong diversion. The Golf series of vehicles exhibits a product line concept that allows auto part variants engine, bodywork, brakes, etc. to be combined as the market requests them. Hence, in terms of profitability you cannot easily compare historic sales and profitability figures from the Beetle. It therefore makes little sense to map historic Beetle to Golf figures. You would instead rather invalidate the value Beetle in the attribute Type of Car and ultimately drop associated business transactions from analysis. Both types of cars would be analyzed separately for their sales, profitability, or any other measure.

The Linnaean biological classification system [16] also features extinct organisms, for example Animalia → Metazoa → Deuterostomia → Chordata → Vertebrata → Mammalia → Placentalia → Carnivora → Fissipedia → Felidae → Machairodontinae (e.g., the saber-tooth tiger).

4.3 Integrate Category with Existing



The environment you operate in is subject to change as well. Markets have an influence on what you can sell. Regulators apply the law of your jurisdiction and determine what your company must report. Political decisions can affect almost any aspect of doing business. The importance of categories—like the type of products or services your company sells, processes, or reports on—ebbs and flows as an effect of these changes. It is therefore not surprising that some categories may fall out of fashion or become operationally irrelevant.

While it is understood that they must be phased out, the question is what to do with existing business facts associated with them. Sometimes an outdated category has a close resemblance to another category that continues to be used, for example because they originated from a joint ancestor. From the perspective of information analysis, it is reasonable in this case to re-categorize the business facts associated with the outdated dimensional value and associate them with the survivor value.

Obviously, this way of re-interpreting history will have side-effects. Depending on how much business was associated with the outdated category (or categories), the survivor experiences a sudden spike of change that cannot be explained unless one remembers the cause. Certainly it has nothing to do with the underlying business reality associated with the survivor. As such, and this is what the pattern highlights, you should have ways of assessing the effects of merging the business of a number of categories into one survivor.

Which survivor do you choose to replace outdated values, how do you time your change and how do you map outdated values?

A loss of business relevance may be temporary. You should wait a certain time to ensure a category is truly outdated and can be considered permanent. This, on the other hand, may worsen the distorting effects of re-categorization.

If you integrate (one or more) categories with the MISCELLANEOUS TERM, you don't have to bother with the question whether the decline in importance is permanent, but that category may grow rather large over time, causing problems of its own.

It is not unusual for business to demand keeping the outdated values in the data stores for quite long. Bringing change propagation to a conclusion can require a lot of time.

Therefore:

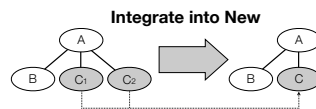
Choose a value that has a close resemblance to the outdated value(s) and remove them from the attribute; change their life cycle status to inactive or superseded; map the old values to the survivor.

If the values are organized hierarchically (as is typical in MOLAP) the parent value of a set of values may take on the role of the survivor. However, that is reasonable only if all of its children lose their importance.

Data stores can detect that they should treat the old value differently at their interfaces. (Derived) measures associated with the old values must be recomputed. Computationally this is less expensive than Integrate Category into New, because only parts of the recorded facts must be re-categorized, but not the survivor. It is no longer intended to record business transactions for the old values, but for the survivor.

When the Federal Republic of Germany and the German Democratic Republic (re-)united in 1990, the former took over the legal and financial rights and duties of the latter. One continued to exist as a political entity, while the other vanished. Hence, from the perspective of the attribute Country, business conducted in the German Democratic Republic was integrated with existing business conducted in the Federal Republic of Germany.

4.4 Integrate Category into New



When certain categories describing the nature of your business become outdated, it may happen that a survivor candidate from out of them cannot be identified. The question may now be raised what to do with the business facts stored in your data stores.

You may want to re-categorize them, in which case a new category must be introduced. The history of business associated with outdated categories is cut off and the dimensional values describing facts mapped to a new one. This is common when you intend to conduct time-series analysis that presents old categories as new ones in the amended view. For that to make sense, the new category must have a close resemblance to the old it bundles together. However, it may happen that there is no point in re-categorization. For example, imagine a car company selling a particular type of car. In July 2003, VW stopped production of the Beetle. From an analytical perspective it would not make a lot of sense to compare the historic performance of sales of Beetle cars with current sales of, say the Golf line of vehicles. In such a situation the facts associated with the Beetle would remain to be so, no mapping would take place.

What is a good choice of successor to replace the outdated values and how do you map them?

Sometimes you will find yourself combining the names of all old values, concatenated, into one, effectively making the new business category the union of all previously existing categories. However, these concepts may seem artificial to business people and not help them understand their data any better.

A loss of business relevance may be temporary. You should wait a certain time to ensure a category is truly outdated. This, on the other hand, may worsen the distorting effects of re-categorization.

If you integrate (one or more) categories with MISCELLANEOUS TERM, you don't have to bother with the question whether the dip in importance is permanent, and that category may grow rather large over time, causing problems of its own. However, for now you'll be fine.

Therefore:

Identify a business term encompassing all outdated terms. Remove the value(s) to be phased out from the attribute and change their life cycle status to inactive or superseded; map old values to the survivor.

If needed, you can use COMPOUND TERM or MISCELLANEOUS TERM for the successor.

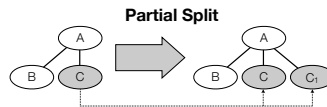
Computationally this is more expensive than INVALIDATE CATEGORY, because all recorded facts must be re-categorized, as there is no survivor. (Derived) measures associated with the old values must be recomputed.

In insurance, so-called run-off business (coverage or portfolios thereof underwritten earlier, but no longer offered to clients) is often detached from normal operations and captured in special vehicles that manage them to optimize cost structures. Run-off business was typically managed earlier in particular parts of the company and then pooled in a different part or even different legal entity. There you will find disparate kinds of coverage that from a content point of view have little in common.

The introduction of the Euro on 1 January 1999 eventually eliminated a host of other national Currencies like the French Franc, Belgian Franc, Deutsche Mark, or Dutch Guilder. For each of these a different conversion rate was set by the European Central Bank, e.g. Belgian Franc 40.3399, Deutsche Mark 1.95583, French Franc 6.55957, and Dutch Guilder 2.20371. The linear factor in here would be the conversion rate, the date of effectiveness 1 January 1999. The change became manifest by means of a SWIFT broadcast on 31 December 1998.

The NAICS [13] industry type 52213 (Credit Unions) was created from the two SIC industry types 6061 (Credit Unions, Federally Chartered) and 6062 (Credit Unions, Not Federally Chartered).

4.5 Partially Split Category



You may find time and again that some part of your business is particularly lucrative or attracts a higher than average transaction volume. In a statistical analysis, transactions would be associated with one particular describing dimensional value more often than others. Your reaction as a business person would be to try to understand what is going on so you can increase profits. One such possibility is that different customer segments purchase your product and that one such segment is interested in a slight variation of the product that appeals more to its needs.

In this case you would want to start offering an alternative product to the identified segment that differs from the existing along one characteristic dimension. The rest of your customer base

would continue to be offered the existing product. Hence, the dimension describing the distinguishing product characteristic would experience a split of an existing category into two. As is the case with INTRODUCE CATEGORY, you have to decide whether to re-categorize historic business transactions.

One further question is to decide whether you really should diversify along (only) one characteristic product dimension. For once, you must make sure that the survivor value retains at least some level of importance. Maybe all of your customers would want to buy a product customized to their needs, effectively eliminating the need for the existing category. Or, which is something even more difficult to handle, the product characteristics appealing to specific customer segments vary along more than one property, which would entail changes to more than one dimension.

How do you choose the new values and how do you map existing data?

Even in a setting where the structure of more than one attribute changes, you may want to use PARTIALLY SPLIT CATEGORY, alas in phases. First, it is sometimes a matter of speculation which value to split and which not. Hence, you may want to try it out one after the other. Second, the data mappings required to re-categorize historic data along a couple of dimensions may become too costly or outright intractable.

In this case, perform splits along each dimension, one by one, treating it as independent of the others. Depending on how late you introduce the new values, transactions conducted previously may not be appropriately categorized. They will either have to be re-categorized, or reported measures will not give a fully accurate picture of history.

With values arranged hierarchically (MOLAP), the survivor can take on the role of parent of the new values with 0% of business facts associated. It is then likely to be re-named and one of its children taking on the parent's old name.

Therefore:

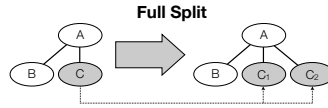
Add the new values to the attribute in the reference data store, life cycle status active, map the old value to the new values.

If the name of the existing category is an instance of COMPOUND TERM, try using parts of its constituent names for naming the new categories. If that doesn't work, try using the existing category as parent of the new ones (i.e., children), in which case you are guided in the choice of their name and meaning.

The new values can be used to record business transactions. (Derived) measures associated with the survivor value and its split-off siblings must be (re-)computed.

The context of this pattern often occurs when organizations are rearranged. For example, a department may have changed its purpose and one team in it should rather be part of a different (new) department. It is then moved over, and the headcount of both the original and the new department changes.

For a while, our company tracked the Line of Business Aviation & Space (covering operation of, e.g. satellites, commercial airplanes, or spaceships) without further distinction. We wanted to understand them at a more fine-grained level, resulting in two lines, both children of the existing, namely Aviation and Space. Previous transactions were split up between the two new so that Aviation & Space did not have any business facts associated with it, anymore.



4.6 Fully Split Category

As opposed to situations where PARTIALLY SPLIT CATEGORY might be more applicable, you may run into a situation where the old (outdated) value has lost all its significance for analytical purposes. Here the main question concerns the choice of successors.

When a value becomes obsolete, you may be able to identify other (new) values that seem appropriate, but do not represent all business transactions to be re-categorized. You may want to continue your search, but that might merely lead to a different composition with a few transactions still not re-categorized. You may be tempted to use MISCELLANEOUS TERM to cover these cases, but this effectively anonymizes them; their business meaning will become unspecific.

It may help to take a look at the name of the business concept represented by the value. This name gives hints as to what concepts it subsumes, especially when it is of the form One Thing & Another Thing. When this concept ceases to exist, the constituents One Thing and Another Thing may continue to have business significance. These are the obvious choice for replacing the outdated value.

How do you choose the new values and how do you map existing data?

With values arranged hierarchically (MOLAP), you will typically want to partially split up an existing value, but retain it with 0% of business facts associated.

Sometimes you may find that you ended up with a number of new values replacing the existing one, but a (small) number of transactions cannot be properly re-categorized, effectively requiring you to add a MISCELLANEOUS TERM.

Quite often you will want to choose new values whose names, concatenated, form the name of the existing value.

Therefore:

Add the new values to the attribute, life cycle status active, and map the old value to the new values. Remove the old value from the attribute and mark it as superseded.

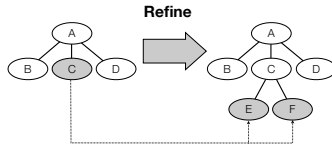
If the name of the existing category is an instance of COMPOUND TERM, try using its constituent names for the concepts behind the new categories. If that doesn't work, try it with less constituents and assign the remainder to MISCELLANEOUS TERM.

The new values can be used to record business transactions. It is no longer possible to record business transactions for the old value. (Derived) measures associated with split-off values must be computed.

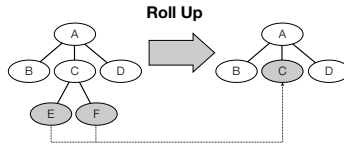
The split of Czechoslovakia into the Czech Republic and Slovakia is a typical case in point. The old Country ceased to exist, and legally the two new ones took over.

4.7 Refine Category

This is a placeholder for a pattern under development. For the problem and solution statement, please take a look at the thumbnails in Table 2.



4.8 Roll Up Categories



This is a placeholder for a pattern under development. For the problem and solution statement, please take a look at the thumbnails in Table 2.

5 Separating Responsibilities in the Application Architecture

Pattern	Problem	Solution
Reference Data Store	What system in your information architecture should be given responsibility to manage the lifecycle of attribute values and store their history?	Store the history of changes to values and their Value Lifecycle in a dedicated repository.
Mapping Data Store	Given the many forms of (temporal) mappings, what should be an optimal scope?	Store the (successor) mappings of values for each measure that their attribute segments, between two points in time. Exposes mappings to other data stores based on the validities and identities of the underlying Reference Data Store.
Value Life Cycle	How do you support the phase-out of attribute values without running too large an operational risk?	Make the value lifecycle at least comprise the set "Active", "Inactive", and "Superseded". Once a value becomes "Inactive," it must become "Superseded" within a timeframe set by the attribute owner. Validate data at inbound interfaces of data stores. "Active" values are always accepted, "Superseded" always rejected. "Inactive" values may be accepted or not, depending on business and technical needs.

Table 3: Thumbnails for patterns helping you to separating responsibilities in the application architecture.

The thumbnails in Table 3 give an overview of the problem and solution of the patterns dealt with in this section, which are not further elaborated at this point.

References

- [1] Francis Anderson: A Collection of History Patterns. Pattern Languages of Program Design 4, Downloaded from <http://www.smallmemory.com/almanac/Anderson99.html> on 27 December 2005
- [2] Massimo Arnoldi, Kent Beck, Markus Bieri, Manfred Lange: Time Travel: A Pattern Language for Values that Change. Downloaded from <http://www.manfred-lange.com/publications/TimeTravel.pdf> on 27 December 2005
- [3] Association for Computing Machinery: Computing Classification System. Downloaded from <http://www.acm.org/class> on 24 March 2005
- [4] Heise Online: Massive Zunahme von Werbe-Spam für gefälschte Luxusartikel (22 March 2005, in German). Downloaded from <http://www.heise.de/newsticker/meldung/57817> on 27 April 2005
- [5] Martin Fowler: Patterns for Things that Change with Time. Downloaded from <http://www.martinfowler.com/ap2/timeNarrative.html> on 27 December 2005
- [6] Ralph Kimball, Margy Ross: The Data Warehouse Toolkit, 2nd Edition: The Complete Guide to Dimensional Modeling. John Wiley & Sons, 2002
- [7] Journal of Economic Literature: Journal of Economic Literature Classification System. Downloaded from http://www.aeaweb.org/journal/jel_class_system.html on 29 December 2006
- [8] Gerard Meszaros, Jim Doble: A Pattern Language for Pattern Writing. In: Pattern Languages of Program Design 4, Addison-Wesley 2002
- [9] Steve Peterson: Stars: A Pattern Language for Query-Optimized Schemas. Pattern Languages of Programs (PLoP) 1994. Downloaded from <http://c2.com/ppr/stars.html> on 7 June 2005
- [10] Andreas Rüping: 2D History. Versioning in the Presence of Retroactive and Future Changes. Pattern Languages of Program (PLoP) 2002. Downloaded from http://www.rueping.info/doc/Andreas_Rueping--2D_History.pdf on 27 April 2005
- [11] Richard Snodgrass: Developing Time-Oriented Database Applications in SQL. Morgan-Kaufmann, 2000. (This book is out of print, but can be downloaded from <http://www.cs.arizona.edu/people/rts/tddbbook.pdf> as of 29 December 2005.)
- [12] Sophos, Plc.: Press Release Sophos PureMessage introduces new spam classification support (12 March 2004). Downloaded from <http://www.sophos.com> on 27 April 2005
- [13] United States Census Bureau: North American Industry Classification System, Revisions for 2002. Downloaded from <http://www.census.gov/epcd/naics02/> on 27 April 2005

-
- [14] Wikipedia: Linnaean Taxonomy. Downloaded from http://www.wikipedia.org/wiki/Linnaean_taxonomy on 29 December 2006
- [15] Wikipedia: Time-Series. Downloaded from <http://en.wikipedia.org/wiki/Time-series> on 10 June 2005
- [16] Wikispecies: Machairodontinae. Downloaded from <http://species.wikipedia.org/wiki/Machairodontinae> on 14 November 2005
- [17] World Health Organization: International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Version for 2006. Downloaded from <http://www.who.int/classifications/apps/icd/icd10online/> on 25 January 2007

Acknowledgements

Thanks go to our shepherds (EuroPLoP 2005: Dirk Riehle, EuroPLoP 2007: Wolfgang Zuser) for their comments on this pattern language as well as the encouragement to take on the issue of time and history in business systems. Thanks go also to Boris Bokowski and Dora Zosso for their comments on earlier versions, specifically illustrative examples and document structure. Finally, thanks go to all participants of Workshop B at EuroPLoP 2005 (Allan Kelly, Lubor Sesera, Halina Kaminski, Tim Wellhausen, and Dirk Schnelle) for their very constructive feedback and, as well, their encouragement for continuing to work on this topic.