TAXONIC

The Taxonomy Revolution

Part I: Knowledge Models

A Taxonic whitepaper, by Jan Voskuil @Taxonic taxonic.com

TAXONIC

Contents

Introduction

Semantics plays an increasingly important role in business processes and IT-systems. This has profound consequences for organizations.

Computer systems can now reason based on the semantic features of a specific case. While the concept of such self-reasoning systems has been proven already in the 1970's, real world solutions now start hitting the market. Consequently, techniques developed in the field of information retrieval and library science are applied to business modeling. Information retrieval, search technology and business process management converge, resulting in revolutionary change. This series of whitepapers explores the backgrounds and the consequences of this change.

Part I addresses the basic concepts: knowledge models, taxonomies, ontologies. Part II explains practical applications of these concepts in the domain of information retrieval and search. Part III is about using knowledge models to support business processes. It discusses real life examples showing how major savings in cost go hand in hand with significant improvement in quality.

The most visionary product of the moment in this area is Be Informed Business Process Platform. Working examples will be presented in this series based on this product. Other examples of products in this field are, for instance, IBM WebSphere Operational Decision Management, Oracle Policy Automation and Isis Papyrus.

1 The Taxonomy Revolution

In his seminal bundle of essays entitled Mastering the Unpredictable, Keith D. Swenson elucidates how adaptive case management will replace previous approaches to supporting business with IT. The traditional paradigm for business automation has been mass-production and the assembly line. It is certainly a good idea to standardize terminology and processes, but for most organizations, this approach has degraded into shoehorning knowledge workers in rigid, uncomfortable patterns. They are forced doing routine work, with countless exceptions to the routine. Unfortunately, IT-systems based on the assembly line metaphor deal badly with these. Workers are constantly busy finding workarounds to deal with these exceptions. This is one of the reasons that knowledge worker productivity is still the biggest of the 21st century management challenges.

Adaptive case management, also called dynamic case management, does not try to fix the paradigm, but rather advocates a new approach from the ground up. Mass-production and the assembly line are metaphors that simply do not apply to knowledge intensive processes. The new paradigm is the TomTom.

IT-systems for supporting knowledge workers must function as a navigation device that helps them navigate a complex maze of optional paths and bring each case to a good end, swiftly and ergonomically.

The fundamental ingredient in such systems is the knowledge model. A knowledge model expresses the rules and concepts that drive a business, and describes regular cases as well as "exceptions" — which then become normal. It forms the basis for deciding what to do with a given case in a given situation. Knowledge models have been with us through the millennia. They are traditionally used in libraries to make knowledge residing in books accessible and help interested parties find their way. This is why findability and business process management now start to converge.

1.1 From the old to the new

In the old world, IT-departments work with procedural models that have only limited expressive power. Take for instance the boxes and arrows in process models. A box stands for an activity, an arrow means "when finished, go to the next". Being generic and abstract, the modeling technique can be applied everywhere, in any domain. But business processes are not about which activity precedes which. They are about executing complex policies, regulations and legislation, about composing products and services tailored to the customer's needs, based on rules that often change, and about finding relevant information from disparate sources — in short, about things that are specific to the given business domain.

Therefore, in the old approach, the procedural model must be accompanied by documents describing the domain specifics. The documents and drawings are handed over to the IT-department to be translated into programming code.

During the testing stage, ambiguities in the requirements are often discovered, leading to changes in the documents, changes in the code, and testing everything over again.

Understand your business

In the new world, business needs are expressed in knowledge models that are transparent to users and business persons. The models can be published to a broad audience. Used this way, the models become an important tool for knowledge management. Decisions made in the board room and those made on the floor are connected. The concepts underpinning the processes in the business and their relations are made explicit and accessible for all, fostering a shared language and understanding.

Cut the development and change process short

Business persons can read and understand the models and can spot wrong interpretations immediately. The resulting knowledge models are directly executable, without translation effort. A change in the model is reflected directly in the documentation and in the way the system operates. This cuts the development and change process short by orders of magnitude.

The models replace specification documents in the old world. When changes in the legislation or policies arise, the models are updated directly. Simone Dobbelaar, CIO of the IND, is quoted saying that implementing a change in legislation used to be a process of months. Now, using knowledge models, it is a matter of days.

Predict the effect of changes

The ease with which changes can be realized has an important side effect. Changes in legislation can be deployed in a simulation environment, and existing cases can be run through the new process. This gives policy makers the opportunity to see what the consequences of different scenario's mean in practice — not only in terms of effects on the outside world, but also the effects on the business processes and effort that goes into them. The effects of policy changes become much more predictable.

Empower knowledge workers

The way knowledge workers interact with business support systems changes fundamentally. Systems based on business knowledge models are far more flexible and user-friendly. Predictable routine work is automated away.

Problem solving work is fully supported. Such systems allow for the necessary amount of freedom needed to deal with the unexpected. Different persons can look at the same case data and documents at the same time, leading to new levels of collaboration. The case worker is allowed freedom to invite specialists at his own discretion, but within bounds.

Evolutionary pressure to change

Adaptive case management technology changes the market. Its core benefits are better life cycle management and better process support.

- Process support applications can be realized faster and are far more resilient to change. Be Informed claims its customers experience 30% reduction of cost of operations, 60% total cost of ownership savings and 90% reduction of the cost of change.

- In addition, organizations can better support customers and knowledge workers. IBM promises 5% growth in customer retention, 30% increase in revenue, and 40% increase in productivity.

As more and more real world examples prove such bold claims to be real, the pressure will rise to make the switch from a procedural approach to knowledge based system engineering and adaptive case management. Reluctance to change will gradually become an obstacle for success.

New skill sets are required to support this switch, however. To optimally benefit, the knowledge models to be used have to meet high quality standards. Creating such models is ultimately a new trade.

1.2 Revolutionary change requires new skills

In 1898, New York hosted the first international city planning conference. The main point on the agenda was horse manure. As Steven Levitt and Stephen Dubner point out in hilarious detail in their SuperFreakonomics, millions of pounds of manure were deposited every day in New York, as was the case in other large cities across the globe. Dung lined the streets and in some places, manure was piled up sixty feet high. This was detrimental to health and posed insurmountable environmental issues. Realistic proposals for a solution were not forthcoming, and the conference was dissolved. And then the problem simply vanished. Not because horses stopped producing manure, but because of a technological innovation: the car. Not that the car solved all problems. It made one set of problems disappear and created a new set: parking, congestions, fuel logistics and so on. At the same time, it raised the level of mobility by orders of magnitude.

Adaptive case management will do the same to IT as what the car did to mobility. Alignment of business and IT has been one of the most central problems in business automation. Adaptive case management redefines the problem by combining advanced execution environments with new modeling languages. In that way, business and IT speak the same language.

Old problems, like tracing application features to requirements through unwieldy sets of specification artifacts will go away. New problems will arise. Most of the work in knowledge based process management is about the quality of knowledge models. The questions that go with that have to do with choice of terms, relating these terms to each other explicitly, and structured sets of criteria for taking decisions. This is the domain of constructing useful thesauri, taxonomies and ontologies — a complex domain that requires special skills.

The difference between procedural versus knowledge based process modeling can be compared to language versus grammar. In theory, you could describe a language simply by enumerating all grammatical sentences of that language. The smarter, practical alternative is that you try specifying these in terms of a finite set of rules: a grammar. A grammar makes implicit knowledge about the language explicit.

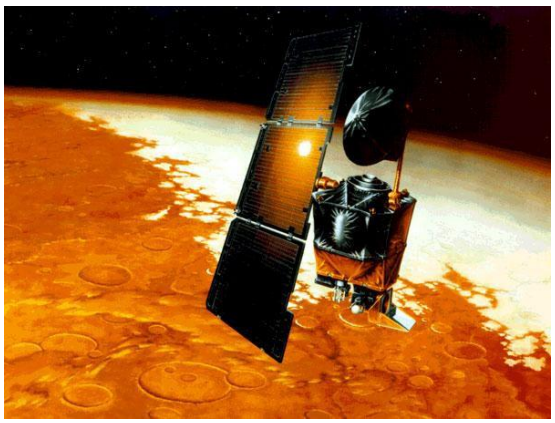
Procedural modeling enumerates all possible permutations of actions in terms of which precedes which, but leaves implicit why some permutations are permissible and others are not. Using a finite set of rules, you can predict the permissible sequences. A knowledge model makes these rules explicit.

A native English speaker knows his grammar, but that is not the same thing as being able to write a high-quality grammar of English. Similarly, being able to run the day-to-day operations of a business does not automatically imply having mastered the art of creating knowledge models from which these operations can be derived.

As we will see in the remainder of this paper, setting up thesauri and taxonomies is subject to subtle rules. It is far from trivial to deliver quality in this area. Like the art of grammar and dictionary writing, this is a specialized skill. It is not something that the traditional programmer can do, either. Part of the skills relate to modeling expert systems, other parts relate to logic analysis and business ontology definition. Adaptive case management puts the business in the driver´s seat, but the IT department and the business need to learn to drive together. After all, driving a car is different from driving a horse.

2 The Aboutness Problem

Soon after the advent of information technology as a separate engineering discipline in the 1960's, it became apparent that semantics is of the essence. It is important to get a computer to do computations correctly and efficiently, but it is just as important that the data you feed into it are about the intended thing. Even within the bounds of a single system this can be challenging.



Figuur 1. The Mars Climate Orbiter (source: Wiki Commons)

Wrong interpretation of essentially correct data is a major source of problems, as famously illustrated by the Mars Climate Orbiter — the spacecraft that crashed because one part of its navigation program interpreted thrust metrics in Newtons and another in Pound-force. In an interconnected world, matters become even worse. How can you be sure your system interprets data correctly when disparate computer systems are connected and data may originate from anywhere?

Similar questions arise in the field of information retrieval. The Web provides oceans of information, which leads to the perennial problem of how one can find information of a particular kind

about a particular topic. The question then becomes, how do you know whether a particular piece of information is relevant? The two questions are in their essence really the same and center round the all-important notion of aboutness. In both cases, the underlying question is about the frame of reference against which a piece of data or a document is interpreted. It is this frame of reference that causes things to have a meaning and to be relevant.

Taxonomies, or more generally, ontologies and semantic networks, constitute a valuable tool for information specialists to deal with such questions. As Heather Hedden points out in The accidental taxonomist, the number of magazine and trade journal articles about taxonomies shows a ten-fold increase in the ten years between 1997 and 2007. This is not surprising.

Rather, it is surprising that the art of managing taxonomies is still not truly a mainstream subject.

Part of the reason for this is that technology providers focus on doing efficient computations rather than on the meaning of data. A suite of products has become available over the years to set up, maintain and apply taxonomies, but these products are expressly oriented towards information retrieval, topic analysis and website publishing. They help organizations to systematize their information products and make them easily available. Important though this is, it is only one side of the aboutness problem. When it comes to running and controlling business processes, semantics is still very much an underdeveloped topic.

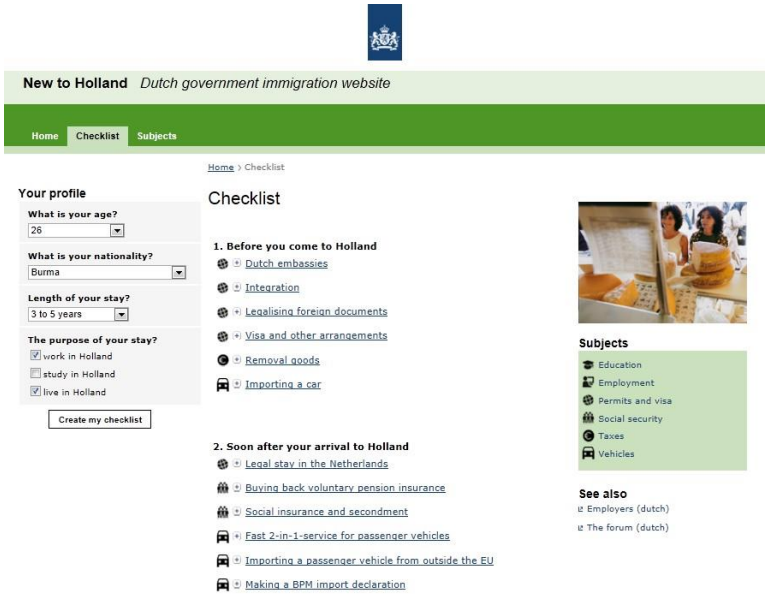
Let us look at some real life examples of how search and business process support converge.

2.1 Knowledge Models for finding relevant information

Dealing with public authorities is not always easy. Citizens find themselves confronted with a large number of different institutions, laws, permits and websites, each offering only a part of the information and services they are looking for. The Dutch government uses semantic technologies to improve its service radically.

One example is a website for foreigners coming to the Netherlands. The website asks questions about the user's context: how long you want to stay, what are the purposes of your visit, and so on. Based on the user's answers, the system assigns a profile, which in turn determines which topics are relevant. The system generates a checklist that sums up information on useful topics, such as legalizing documents, importing a car, making proper arrangements for social security, and so on.

The information assets displayed in the checklist are provided by a large set of public organization, such as IND (the immigrations office), the tax department, UWV (benefits payment), RDW (vehicle registration), and so on. Each is content provider is responsible for managing his information assets. This includes associating metadata to each asset, using a shared, structured and controlled vocabulary of indexing terms.



Figuur 2. Website newtoholland.nl

Thus, if your age is 71, your checklist will contain topics such as Old Age Pension. If your age is 16, you are not entitled to drive a car, and no topics will be included related to driving. This leads to a citizen centric approach to publishing governmental information. The citizen is just interested in the information relevant to his situation, and does not care which of the 1300 Dutch agencies provide it. Selecting the information that is relevant given your specific situation is a knowledge intensive task.

A set of knowledge models is used to determine the user profile based on the user's answers, and define a mapping from profiles to relevant topics. These topics are then matched against metadata associated to individual information assets. Thus,

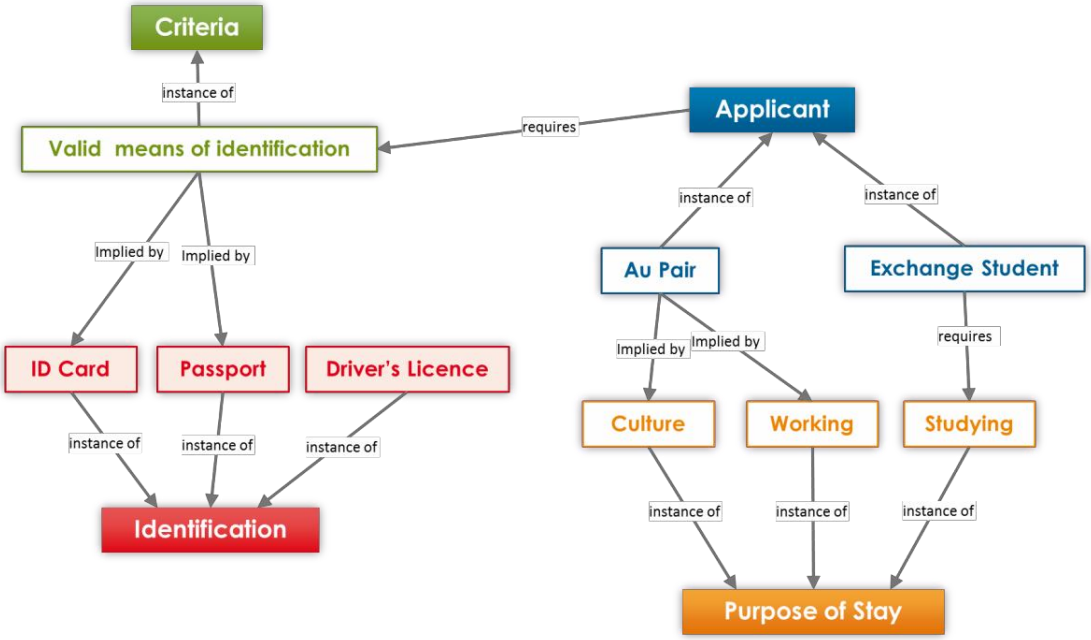
in order to determine whether particular pieces of information are relevant to a user, two steps are taken:

- Knowledge models are used to reason about the user's context and derive a specific query in terms of a controlled vocabulary.
- Information assets are tagged with metadata from the controlled vocabulary defined by the knowledge model.

We will discuss this example in more detail in Part II, but the important point here is simple: knowledge models are an essential ingredient in supporting search applications.

2.2 Knowledge Models for running business processes

Let us now turn to an example of knowledge models used to support business processes. The IND, the Dutch immigration office, uses similar knowledge models to assign profiles to cases. Thus, for each application for a permit to visit the Netherlands, it can be determined which rules are relevant for that particular case. The following figure shows a small part of these knowledge models.



Figuur 3. Knowledge model showing immigration concepts

This model expresses a number of business rules, for instance that an Applicant of type Exchange Student must have Studying as purpose of stay, and that being an Applicant of any type requires a valid means of identification. These rules determine how the process for each case proceeds. They determine the status of an application for a visa, the next steps that are possible given the status, the options that the IND worker has while working on the case at a particular point in time, and so on. By describing the application and its context in such meaningful terms, it becomes possible to make inferences automatically. The execution environment directly executes these models.

Strikingly, the knowledge models in both examples — the website newtoholland.nl and the process automation at the IND — are realized using the same technology, in this case Be Informed Business Process Platform. They use the same modeling language and are structured in the same way. In fact, the knowledge model in figure 3 above can be shared across both systems. This shows that the domain of case management and the search domain are indeed converging. In the remainder of this part of the series we discuss the nature of these models in more detail, independent of technology implementing them.

3 Taxonomies make the world understandable

Let us start this guided tour on knowledge models in business contexts — or business semantics for short — by reflecting on why there is apparently a deep seated urge to develop and use taxonomies.

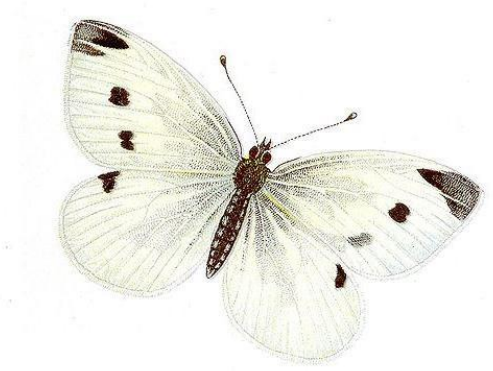
3.1 The oldest profession in the world

The aboutness problem is a problem of meaning. A systematic way of assigning names to things in a given domain is a great help in this regard, because it structures the way we communicate and reason about the domain. It is no coincidence that many consider taxonomy construction the world's oldest profession. Identifying edible plants is of critical importance to hunter- gatherers, and being able to communicate about these carries tremendous evolutionary advantage. In any case, written attempts to systematize knowledge of plants have been documented as early as 3000 BC in China.

Such attempts find difficulty with some aspects of natural language. We use language to do a variety of things: motivate people, solicit for help, develop new ideas, forge social bonds, perform rituals, and write poems — to mention just a few. Language is a multipurpose tool with unprecedented flexibility.

This flexibility sets human language clearly apart from languages in the animal world, but poses challenges when it comes to pure and clean information exchange. As Anna Pavort points out in her fascinating history of naming plants, The Naming of Names, a widespread flower such as the marsh marigold (Caltha palustris) has about sixty common names in France, another eighty in Britain, and at least 140 in Germany, Austria and Switzerland. With the rise of modern science in the seventeenth and eighteenth century, many initiatives were taken to constrain the freedom of language.

The most famous propagator of this movement is no doubt the legendary 18th century taxonomist Carl Linnaeus, who invented, or rather perfected, a carefully designed naming scheme for living creatures, based on a shared taxonomy and explicitly defined norms for applying a given name to a given creature. Based in this naming scheme, the Small White now has the scientific name Pieris rapae, and naturalists can be confident that they are talking about the same kind of animal when discussing this remarkable little butterfly. As we will see, there are still many valuable lessons to be drawn from the method that Linnaeus developed.



Figuur 4. Pieris rapae (source: Wiki Commons)

While flexibility is necessary for creativity and art, science demands more rigidity. After Linnaeus' approach to the taxonomy of living things became accepted in the field of biology, other fields soon followed. Naming conventions and grouping principles were developed for chemistry, medicine, astronomy, geology, mineralogy and so on. Also in law and business naming conventions are

increasingly important. Think for instance of the RAL system of naming colors, used in the paint industry.

3.2 Adequate naming systems

An adequate system of naming has three important properties:

- Standardized. Everybody uses the same term for the same thing.
- Structured. There are clearly defined relations between the different names. For instance, the butterfly species called Pieris rapae belongs to a family of butterfly species, and this family is named Pieris.
- Systematic. There is a clearly defined method for determining which name applies to a given instance or object.

Let us reflect on these three properties. The need for standardized names is the most obvious of the three. Standardization of naming is an ongoing concern in countless domains, inside many organizations, across and beyond. Conflicts can be difficult to resolve due to political or practical considerations. However, the most essential precondition for a standardization effort to be succesful is making the system accessible. An easily accessible thesaurus that helps users to quickly find the right term is of great help for this purpose.

A useful system of naming also adds some structure to a domain. The most well-known structure is the taxonomy, a hierarchical ordering of class names, or taxa (singular taxon). The relation between a taxon and its parent is usually some sort of subclass relation, so that one can infer that each mammal is also a vertebrate. Conversely, if someone talks about a vertebrate animal, you know that the animal referred to is certainly not a butterfly. The importance of being able to make such inferences can hardly be underestimated. Structural errors in a taxonomy that prevent this type of reasoning, effectively render the taxonomy useless. We will come back to this later on in this whitepaper.

The third property of a naming system is that it defines a method for applying names with clear cut criteria. This aspect is often forgotten. In the field of information retrieval and enterprise content management, there is an extensive literature on how to categorise information using different kinds of structure. For instance, one can decide that in a thesaurus the term "football" is to be used as a narrower term of "sports". But when do we decide to apply the label "football" to a book or article? Contrary to appearances, the answer is not always obvious. In practice, subject experts are hired to catalogue book collections, but the precise way these experts reason is most often left implicit.

Linnaeus, on the other hand, came up with very precise criteria that allow one to decide which species a given plant or animal belongs to. He famously devised a way of grouping plants based on the number and arrangement of petals, which he described as "serving as the bridal bed which the great Creator has so gloriously prepared in order that the bridegroom and bride may therein celebrate their nuptials." This was deemed deeply shocking, especially for women, and created outrage in religious circles. Worse, the system did not lead to useful insights in plants. For this latter reason, Linnaeus' "methodus plantarum sexualis" was soon replaced by another system.

In the next section, we will touch on the reasons for preferring one grouping method over another. At this point, it is important to see how important naming

and grouping criteria are. The systematicity of a naming system is in a sense its most important property. When you think about it, a naming system cannot be standardised and structured without clear systematic criteria.

3.3 Single-access keys

When these systematic criteria themselves are grouped in a useful way, we arrive at a "key", also called single-access key in biology. It is useful to consider this in some more detail, because the principle also applies to knowledge models in adaptive case management, as we will see in Part III of this series.

Such a single-access key, as used in biology, is a set of logical choices that together constitute a decision process. The user of the key is led through a number of questions about the object to classify. At the end of the process, the object is fully classified. The following is the first part of an example of a single-access key in the English Wikipedia. It classifies United States oaks:

1. Leaves usually without teeth or lobes ->2; leaves usually with teeth or lobes ->5
2. Leaves evergreen ->3; leaves not evergreen ->4
3. Mature plant large tree — Southern live oak Quercus phellos

   Mature plant small shrub — Dwarf live oak Quercus minima
4. Leaf narrow, about 4-6 times as long as broad — Willow oak

   Quercus phellos

   Leaf broad, about 2-3 times as long as broad — Shingle oak Quercus imbricaria
5. (Etcetera)

So if you stand before an oak tree, somewhere in the US, and its leaves are without teeth or lobes, and the leaves are evergreen, and the tree is quite a large tree, then you know you stand before a Southern live oak.

Such keys are often published in the form of field guides. Sales figures show that such guides are quite popular, in different fields: birds, trees, flowers, and so on. As noted earlier, it is part of human nature to consider it important and fun to know names of things around us.

There are different kinds of keys. A key can follow the taxonomy very closely which provides the underlying classification. In such keys, there is a classification question for each node in the taxonomy. You navigate the taxonomy from the top down until you reach a leaf.

Alternatively, the key takes shortcuts. It limits the choice of characteristics to those most reliable, convenient, and available under certain conditions. A field guide to songbirds is an example of such a key. Scientific keys of the former type are called synoptic, the practical keys of the latter type, diagnostic keys.

So we find ourselves in a complex situation. Taxonomies may be replaced by other taxonomies, such as happened to the one Linnaeus proposed. At the same time, for any given taxonomy, different keys may coexist, depending on the needs of the intended audience. This insight is quintessential for modelling business processes,

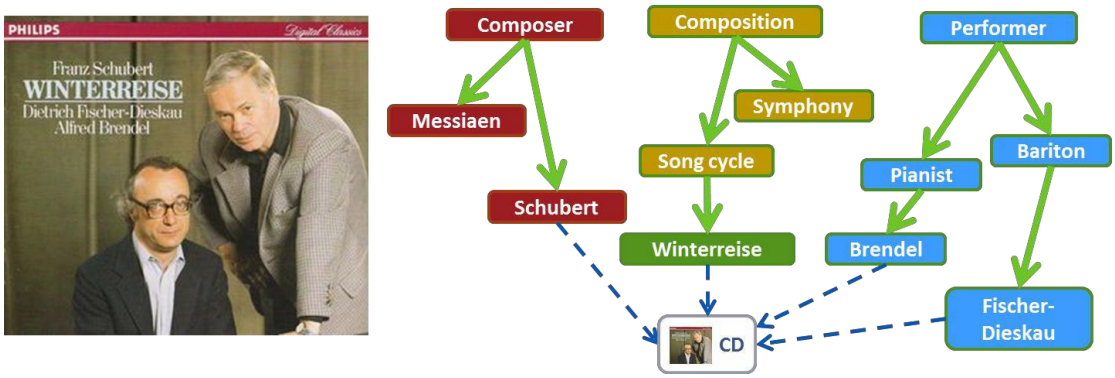
as we will see below. First, however, we take a closer look at what taxonomies and ontologies really are.

4 What are taxonomies and ontologies?

4.1 Some notes on terminology

Introducing terms and definitions works best in the context of an example. A classic application of taxonomies is metatagging, or the art of assigning metadata to objects to make them retrievable. The following example uses three mini-taxonomies, one for classifying kinds of composition, one for classifying kinds of performers, and one to classify composers, to add metadata to a CD. The green arrows mean "a is an abstraction of b". The blue dotted arrows don't belong to the taxonomies proper but rather indicate a relation between taxonomies. They mean "a performs or has composed b".

Taken together, the construction of labels and arrows can be taken as a description of the CD on the left:



Figuur 5. Using taxonomies for metatagging a CD

Such networks of combined taxonomies, when used to provide interrelated metadata for information items, tremendously help information retrieval. The CD above would be one of the search results if you look for CDs in which Brendel plays a song cycle, or that contain works by Schubert, or that contain works performed by just a pianist and a baritone, or on which Fischer-Dieskau performs as a baritone and not as a director, and so on. Such highly specific searches cannot be performed with a generic search engine that just looks at occurrences of search terms in texts. Rich metadata associated to information items, in short, metatags, offers end-users much more powerful ways of finding information than the classical search engine.

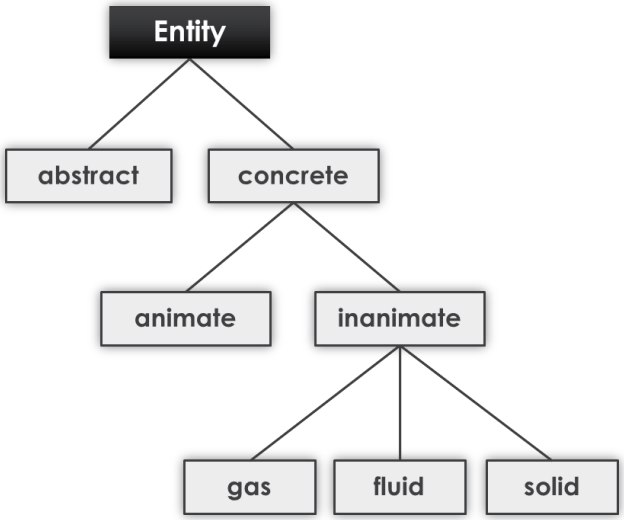
In practice, taxonomies that are linked to each other in this way exemplified above are often called an ontology. Mathematically, both taxonomies and ontologies are graphs, consisting of nodes and directed edges linking them. Taxonomies are strictly hierarchical and impose a partial ordering on the nodes contained in them, while ontologies do not have this restriction. It is in this restricted sense that we will use these terms in the present paper.

The terms semantic network and knowledge model are used interchangeably to indicate networks of concepts. Taxonomies and ontologies as just defined can thus be called semantic networks or knowledge models.

4.2 Speak the language of the business

A brief digression is in order about how we will not use the term ontology. The term derives from Greek philosophy and stands in its original meaning for a theory of

what exists: life, the universe and everything. Such theories are of a metaphysical nature. Most contemporary analytical philosophers agree that such theories don't go anywhere. The problems they pose are either empirical by nature and should be solved by scientists in the lab rather than by philosophers. Or they are really meaningless and, hence, non-problems. Consider the following classical example of a would-be ontology.



Figuur 6. A metaphysical ontology

Many obvious examples of "entities" fit this scheme perfectly, such as Gouda cheese ("solid"), nascent stars ("gas") and poetry ("abstract"). But what about a wrist watch? It may look like a concrete, inanimate solid, but you can't melt it and then make it a solid watch again by lowering the temperature. Is a neutrino a solid? And what about an ant colony? Could you say that an ant colony, certainly not a gas, fluid or solid, is a living thing so that it fits under "animate"? Or are ant colonies abstract like poetry? And what shall we do with nothingness?

These are precisely the artificial, metaphysical questions about which one could think indefinitely, without ever arriving at a useful conclusion. As the great 20th century philosopher Ludwig Wittgenstein once put it, the metaphysicist resembles a child that scribbles something on paper and then asks his dad what it means. In as far as questions of existence are valid, solving them should be left to scientists in the lab.

The desire to formulate an unequivocal, general theory of existence often leads to a naïve perception of language, in which only its representational aspect is taken seriously and all other aspects are disregarded. In reality, use of language can often be understood as a kind of interactive game. Like moving a bishop to E5 in a chess game, uttering a word or a sentence is part of a certain repertoire of behavior. When a surgeon says "scalpel" during an operation and holds his hand up, this is a move which the assistant answers by putting a scalpel forward. The fact that the word scalpel "stands for" a scalpel is not the most important in such circumstances — the sequence of actions it implies is far more significant.

Information specialists should follow Wittgenstein's advice and not talk about what "really" exists and what does not. Rather, they should analyze the language and the behavior of their clients. Language and behavior are intimately connected. To

speak the language of the business is to understand its behavior and vice versa. This is a fundamental insight.

Let us now return to the term ontology in its more restricted sense: a graph of labels, not strictly hierarchical, that can be applied to things in a domain. As discussed in the previous section, this is useful because it structures the way we think and communicate about a domain. But why is that so useful and important? This question can be answered at different levels.

At the deepest level, the structure you impose on a domain may have descriptive, prescriptive, or explanatory purposes. In a business environment, most models are prescriptive or descriptive. On a more operational level, these models can have the specific purpose of supporting information retrieval or rather supporting business processes. The quality aspects that determine whether a model is optimally fit for its purpose depend on these distinctions. So let us discuss this topic in more detail.

4.3 Describe, Prescribe, or Explain

An explanatory model makes predictions about a domain. If these predictions are confirmed, the model helps us understand it. This is what empirical science is about. This idea was made famous by Karl Popper, the great philosopher of science and methodology. Models that make correct predictions reveal something about underlying causal mechanisms. These mechanisms are always hypothetical. When a model makes wrong predictions, the model is falsified, but when the predictions are correct, you still cannot be sure. Maybe a better, as of yet unknown model exists that makes the same predictions based on other mechanisms, and is superior because it makes other additional predictions. This happened when Einstein's relativity theory replaced Newton's model of gravity.

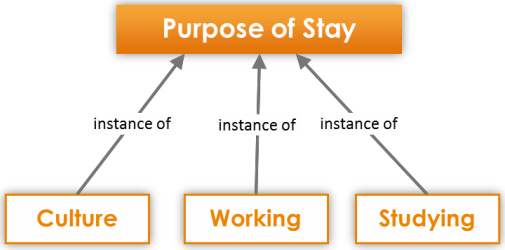
Taxonomies in biology are typically of an explanatory nature. Linneaus' taxonomy of plants, discussed before, was based on the wrong features and failed therefore to lead to insights. Modern biologists have come to understand the underlying mechanisms that cause different species to have different properties. They look at the DNA of organisms and derive the taxonomy from that. A biological taxonomy not only describes a domain, but is also a model of how the different species arose through time. Man and mice share an ancestor that was not the ancestor of the Small White. Should that claim appear false, then the model is falsified, hence wrong.

A descriptive model has no other ambition than to describe a domain. It does not make predictions and cannot, therefore, be falsified in that sense. The only sense in which a descriptive model can be wrong is by providing factually incorrect descriptions. But within these bounds, essentially anything goes. For instance, there is nothing wrong with a taxonomy that only distinguishes hairy and hairless animals. In the worst case, that particular descriptive taxonomy may be not so useful. Most taxonomies used in the search domain are of the descriptive type. Suppose I go to a library and want to look for literature on WordFeud, do I look in the Sports, the Internet or in the Games category?

Neither categorization is right or wrong per se, as long as we all agree to speak the same language.

Prescriptive models are models used to prescribe things. They often arise from legislation or policies. For instance, the norms for handling requests for permits

must be observed closely. The following taxonomy effectively tells us that, as far as the Dutch immigration office is concerned, there are three possible answers to the question as to your purpose of stay:



Figuur 7. Taxonomy of concepts related to purpose of stay

Taxonomies in such prescriptive models can be used to categorize requests, administrative tasks and activities, process steps, process outcomes and more.

Normative taxonomies (and ontologies) are often based on prescriptive documentation such as policies or legislation. Often, the resulting models are more explicit than the underlying documentation. That is usually a good thing, since many sources of ambiguity are then resolved. It also means that the processes of creating such models may pose hairy difficulties. We will discuss modeling legislation in more depth in Part III.

4.4 Taxonomies are everywhere

The differences between explanatory, descriptive and prescriptive taxonomies are often blurred, and grey areas exist. This is amplified by the circumstance that we use taxonomies almost all the time, even if we don´t realize we do.

In fact, taxonomies are pervasive in the very semantics of our everyday language. Different linguistic and cultural groups use different taxonomies to classify the same things. Germanic languages tend to distinguish people more by sex than by age, while Austronesian languages do it the other way around. In Indonesian, there are words for older and younger sibling (kakak and adik, respectively), but there is no single word for the concept of "male sibling".

Javanese lumps together poultry and fish (iwak) as opposed to meat (daging).

Such differences often point at cultural differences. Some cognitive scientists go further and claim that these even shape human cognition. Eskimos have been claimed by some to be psychologically different based on the fact the Eskimo languages allegedly have dozens of words for snow. This is a myth however, hilariously debunked in Geoffry Pullum's "The Great Eskimo Vocabulary Hoax." Yet, there is evidence that language trains thought to at least some extent.

Be this as it may, the importance of taxonomies can hardly be overestimated, even or maybe especially so in a business context. One must keep in mind, however, that the manner in which a taxonomy can be right or wrong depends on its purposes.

5 What makes a taxonomy a good taxonomy?

To better understand how taxonomies can be put to use, this section discusses some basic rules of engagement. There are three essential quality attributes that determine the viability of a taxonomy. The rules governing these are fairly fundamental. They are:

- Distinctiveness. The classes defined by the taxonomy must be distinctive vis-à-vis other classes.
- Uniqueness. Each class in a taxonomy must occur only once.
- Homogeneity. The subclasses of a taxon must be defined so that they all differ in the same aspect.

We discuss these rules in turn.

5.1 Distinctiveness

The rule of distinctiveness is sometimes, in the context of taxonomies, also called taxonicity. The groups distinguished by a taxonomy must be as highly "taxonic" as possible. As Plato put it, a good taxonomy "… divides things in classes precisely where the natural joints are, not breaking any apart after the manner of a bad carver." A number of properties help in this regard.

- A well-known causal mechanism exists that explains how a class differs from the rest.
- There are clear-cut positive criteria for class membership.
- The taxonomy exhaustively classifies a domain.

Causality is a must-have property for taxonomies with explanatory ambitions. For descriptive and prescriptive taxonomies this is not necessary, but having this property is always a good thing. Distinguishing by causes is stronger than by effects. For instance, a taxon defined as people having a high alcohol percentage in their blood is stronger than one defined as people driving like a drunk person. A taxonomy of illnesses better classifies patients than a taxonomy of symptoms.

Positive criteria for class membership also help. People having type 2 diabetes form a clearly defined class, but there is no taxonic class of people not having this disease. Ideally, a taxonomy has positively defined classes that exhaustively classify a domain. Take for instance sex: classifying people in male and female covers all people (with only a few exceptions). Both classes can be defined in terms of clear-cut positive criteria, and the causal mechanism explaining the difference between the two is well-known.

Of course, in many cases it is not possible to arrive at such a high level of distinctiveness. In the days of Linnaeus, nothing was known about what causes the different species to be different. Many attempts of arriving at a well-motivated taxonomy failed — as was the case with Linnaeus' own proposal, as discussed before. The discovery of DNA caused the taxonomy of animals and especially plants to be fundamentally revised. For animals long extinct, such as dinosaurs, DNA is mostly not available, so that for those species the taxonomy has a highly hypothetical character, and hence a high degree of taxonicity is not attainable.

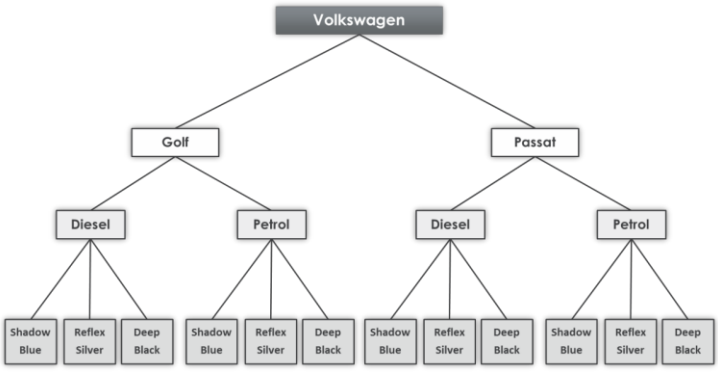
For descriptive and prescriptive taxonomies, distinctiveness is most often less of an issue, as long as they live up to their purpose. We may want to subclassify blog posts in the sports category into posts about football, tennis and other sports, and

that is just fine. However, it is still a good idea to at least strive at an exhaustive classification in terms of positive criteria.

5.2 Uniqueness

The second important quality attribute of a taxonomy is that each taxon is defined in terms that depend in some sense on its parent. More specifically, this rule requires that the criteria distinguishing different subclasses inside a taxon are unique to that taxon. In other words, these criteria should not be reused elsewhere in the taxonomy.

In practical terms, this translates to: every taxon only occurs once. This is extremely important. Consider the following example of taxonomy of Volkswagen cars that can be purchased at a hypothetical vendor. The vendor wants to distinguish all cars on offer in one taxonomy. As you can see, the taxa "diesel" and "petrol" occur twice each, while the three colors each occur four times. This must be avoided at all times, for practical as well as conceptual reasons.



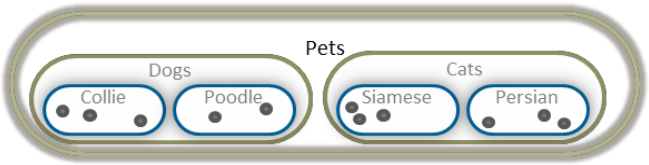
Figuur 8. Example of a bad taxonomy

One can see that such multiple occurrences of a taxon can easily lead to combinatorial explosions. Add a few more car types, and a few more colors, and tree will grow exponentially. Taxonomies change over time due to dynamics inherent in the domain it describes. The cost of maintaining a taxonomy is proportional to its size. Hence, smaller means better.

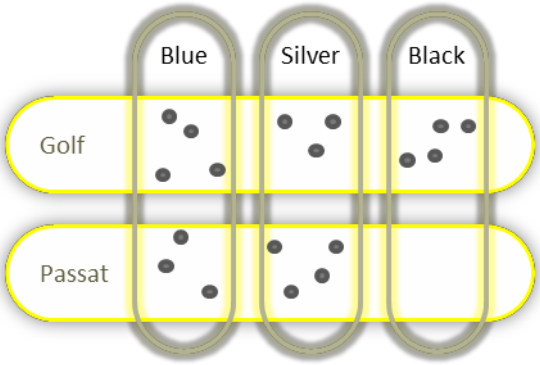
At the conceptual level, however, there is something positively wrong with the above taxonomy. It misses an important point, namely that the defining criteria of the classes involved are apparently completely independent of and unrelated to the criteria that define their parent class. In this example it is obvious that the color of a car is independent of its engine type, and that engine type is independent of the model of the car.

The dependency relevant in this context is that of logical implication. Being member of a taxon logically implies being member of the parent of that taxon. For instance, being a mammal logically entails being a vertebrate. A taxonomy that does not allow one to make such implications is effectively useless. In the taxonomy in figure 8, such deductions do not hold: if a car is member of the taxon Diesel, you cannot conclude that it is a Passat or that it is a Golf — it can be either. It is completely unclear, therefore, what the lines connecting the boxes mean. In any case, they do not indicate parenthood in a taxonomic sense.

As noted, the underlying problem is that engine type and car color classify the car domain in orthogonal subsets. This can be made more precise using basics of set theory. In a proper taxonomy, each taxon is a proper subset of its parent, recursively. The set of poodles is a proper subset of dogs, which is a proper subset of pets. The venn-diagrams in figure 9 visualise this. The venn- diagrams in figure 10 show a different picture: the set of silver cars overlaps with both passats and golfs.



Figuur 9. Properly nested subsets



Figuur 10. Overlapping sets

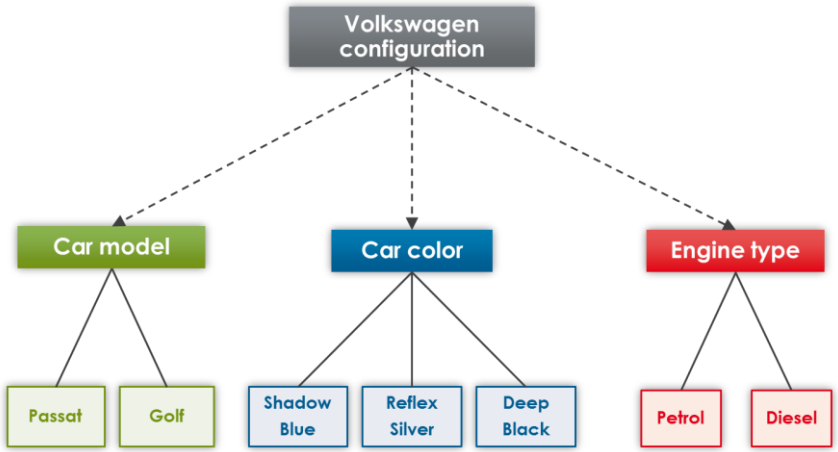
A set can be defined by it extension, which is an enumeration of its elements at a given point in time. A set can also be defined by its intension, that is, in terms of its distinguishing criteria. The set of yellow cars and the set of sports cars may have the same extension at a second hand car dealer — when the only yellow car and the only sports car available happen to be the same.

These two sets, having the same extension, clearly have different intensions. Their membership will change over time, their intension will not. In a taxonomy, each taxon ideally has an intension based on its parent's defining criteria made more stringent by additional criteria that are not used elsewhere in the taxonomy.

In taxonomies with explanatory ambitions, the uniqueness constrained is often an automatic consequence of causality. Subclassification of a taxon in biology is based on distinguishing DNA features unique to that taxon caused by a single (most often hypothetical) mutation. In descriptive and normative taxonomies, such as in the above example, one does not need or even want to be explicit about such underlying causes. Hence, extra attention must be paid to the uniqueness constraint.

This leaves us with one final question: how can we deal with the situation our unfortunate car dealer finds himself in? How can we set up a valid taxonomy to

classify the Volkswagens? The following diagram captures the intended semantics in a much better, more transparent way:



Figuur 11. Refactored taxonomy

Here, we have three taxonomies each classifying independent car aspects: car model, engine type and car color. The dashed arrows are meant here to indicate that the concept of "Volkswagen configuration" requires a value for each of these aspects. Essentially, we have refactored the original taxonomy for classifying cars in three independent taxonomies classifying property values of cars.

It is interesting to think about what would happen if one of the colors, say, Deep black, is not available for all car models. Suppose a Passat can be delivered only in blue and silver, not in black. From the viewpoint of pure information retrieval, there is no need to make such dependencies explicit: in a search application, the search for items metatagged as {Car model: Passat; Car color: Deep Black} would simply return no results.

However, if you want to use these taxonomies in a web shop to support users to actually order a Volkswagen, it is quintessential that we can express such rules explicitly. Otherwise, users can order a car that violates the norms prescribed for the buying process. We come back to the question how that can be done in Part III of this series.

Let us conclude that intensions of classes must occur only once in a taxonomy. This is important to keep the taxonomy maintainable. Moreover, multiple occurrences of a class in a taxonomy indicate a semantic error. This error can be fixed by separating out distinct taxonomies classifying conceptually independent aspects.

5.3 Homogeneity

An important rule of classification theory is that the subclassification of a taxon must be based on a single criterion or aspect. Thus, a taxonomy that classifies cars into "Petrol", "Diesel", "Sports car" and "Minivan" violates this rule, because two separate aspects are used in defining the subclasses: engine type and model type.

Of the three rules — distinctiveness, uniqueness and homogeneity — the homogeneity rule is the weakest and most often violated in everyday life. In Dutch vacation booking websites (as opposed to, say, American ones), it is common

practice to offer three main categories of vacation: flight vacations, car vacations, and ski vacations. With a flight vacation, you fly to your destination, with a car vacation you get there by car. Most of us would be unpleasantly surprised, however, if we were told to ski to wherever we booked our ski vacation to. So this taxonomy violates the homogeneity criterion: the first two subcategories are distinguished by means of transport to the destination, the third, by means of the activity typically performed at the destination.

The problem with this is that certain searches are difficult to perform. Where do I go if I am looking for a ski vacation, and I want to take the plane to save time? This can lead to excessively frustrated users when the item searched for is not there. Unless measures are taken, the prospective vacation booker will be made to browse through pages of flight vacations and then pages of ski vacations, only to conclude there is no offering that fits his requirements.

Peter Becker and others give a number of real life examples of this in their textbook on library science, Organiseer je informatie (2010).

The vacation booking website example not only illustrates the importance of homogeneity in taxonomies, but also reveals something about the importance of cultural habits. If nine out of ten booking sites distinguish the non- homogeneous categories flight, car and ski vacations, then you can be pretty sure that the intended audience — in this case, Dutch vacation seekers — have become accustomed to this way of conceptualizing their vacation. When designing the structure of a new booking site, you may want to observe this custom and deliberately follow the same subcategorization. The question then becomes: how can you speak the language of the business and still avoid problems of the kind discussed above? In the next section we will discuss some of the measures you can take to this end.

5.4 Relations between the rules

Before ending this subsection, let us briefly consider the three rules — Distinctiveness, Uniqueness and Homogeneity — in relation to each other, and also discuss some other well-formedness rules for taxonomies.

A well-chosen causal substrate may automatically imply obedience to the three rules in one fell swoop. We have mentioned the single-mutation hypothesis in biology above. Another example is phylogeny of languages, in which hypothesized mutation is also the basic causal mechanism that explains or predicts the phylogenetic tree of languages.

In descriptive and normative taxonomies, which are typically used in a business context, such an underlying mechanism is by definition not available. In those cases, it becomes clear that the three rules are really independent.

The Volkswagen taxonomy, which we had to refactor because it violates Uniqueness, is perfectly distinctive: the set of cars that are of model Passat and have a diesel engine and are of color Silver does not overlap with any other set and is as distinctive as can be. It poses no problems for Homogeneity either, since each taxon subdivides into subclasses using a single distinguishing feature.

A taxonomy that violates Homogeneity often violates the other two rules — though not necessarily so. The Dutch vacation booking taxonomy that violates Homogeneity does observe Uniqueness. The taxonomy does violate Distinctiveness

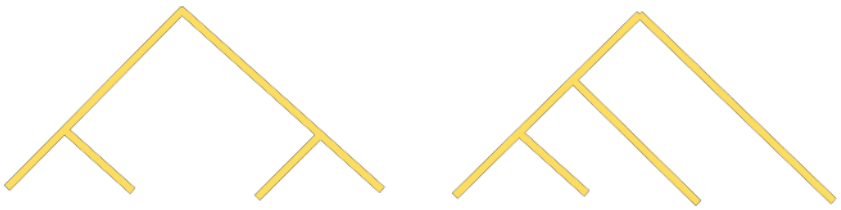
however: flight vacations and ski vacations overlap and are, hence, not distinctive. It is not difficult, however, to come up with examples that are not homogeneous and still in line with Distinctiveness.

In sum, the three rules for taxonomy construction are essentially independent, but they interact in intricate ways. To be able to pry them apart and see how they work is of essential importance to taxonomy evaluation.

5.5 Other rules and non-rules for taxonomies

Two further rules deserve mention in this section: binary branching and symmetry. To start with binary branching, this is most often relevant in the context of explanatory taxonomies. Whenever a taxon has more than two subclasses, there is likely to be something waiting to be uncovered. For descriptive and normative taxonomies, this rule of thumb seldom applies.

A rule widely cited in the literature, to be referred as Symmetry in this paper, holds that a taxonomical tree must be as symmetrical as possible. Each element in the population to be classified should as much as possible have the same number of ancestors. This rule is sometimes formulated in terms of gradation or modulation, so that navigating the taxonomic tree from top to bottom one gradually moves from abstract to concrete, without discontinuities. Thus, according to this rule, the taxonomy on the left is preferable to the one on the right.



Figuur 12. Symmetry in trees

Both trees distinguish four species of elements, but in the left one every element has exactly two ancestors, while in the right one, the number of ancestors varies from three to one.

In biology, asymmetry is the rule, rather than the exception. Species that went extinct 250 million years ago have significantly less ancestors that current species. In some lineages mutations occur more frequently than in others. Pruning and grafting the taxonomic tree of life, which happens regularly as new insights in DNA-sequence develop, most often adds even more asymmetry. Birds used to be a separate class within the phylum of vertebrae, but have recently been "demoted" to form a group within reptiles, alongside the crocodiles.

Also in the context of descriptive taxonomies for purposes of, say, information retrieval, this rule can hardly be said to foster quality, despite the fact that this is often thought to be the case. One textbook example goes like this: you should not place tennis socks directly under the class garments. Rather, you should organize matters as follows: garments > sport´s wear > tennis wear > tennis socks. The question is, however, why?

Of course, when you design a web shop for garments that also sells different kinds of socks you have to speak the language of your customers. Ideally, you would be

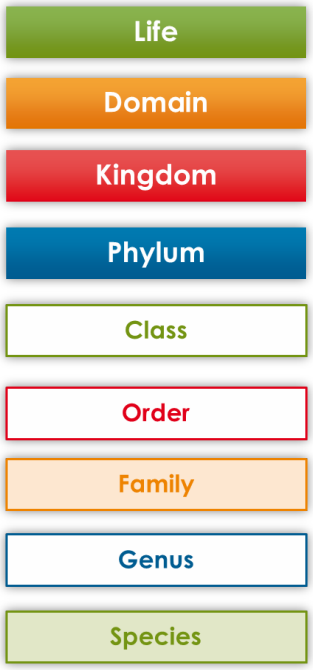
able to employ different taxonomies alongside each other — a garment type taxonomy (shirts, trousers, socks), a taxonomy of circumstances under which to wear the garment (sport, leisure, formal), and so on. But that is a completely different.

A variant of the symmetry rule is a notion that subcategories in a taxon must be equally abstract. Clay Shirkey, prolific author on findability and related subjects, discusses the following remarkable fact in one of his articles. The Library of Congress' categorization of the subject History includes the following subtopics:

- DR: Balkan Peninsula
- DS: Asia
- DT: Africa

However, the Balkan Peninsula is tiny, not to say extremely small, when compared to Asia or Africa, which are continents in themselves. The categorization therefore seems intuitively imbalanced.

The reason behind this apparent asymmetry is quite simply that the Library of Congress happens to have more books on the history of the Balkan Peninsula available than on that of Asia and Africa, at least relatively speaking. Shirkey then goes on to argue that the categorization scheme is a response to physical constraints on storage, and to people's inability to keep the location of more than a few hundred things in their mind at once. In an electronic world, such constraints do not exist anymore, so the argument goes.



Figuur 13. Linnaean ranks

However, even in the electronic world the same constraint actually does exist. The constraint is given by the assets you want to classify and the distinctions that you want to make. Suppose you sell vacations and you have five on offer with destinations in Asia and one hundred and fifty in the Balkan. If that were the case, you probably want to subdivide the Balkan vacations much more often that the US vacations. And it doesn't matter whether you sell your vacations on a website or in an old fashioned physical shop with physical, paper vacation catalogues. The imbalance in the taxonomy follows from the relevant distinction — destination — and the number of assets in each of the classes defined by this distinction.

The symmetry rule has its roots ultimately in metaphysics and the vague notion that abstractness is somehow measurable. The traditional Linnaean taxonomy postulates nine predefined levels of abstractness or ranks in the tree of life, as shown in figure 13. Modern phylogenetics and cladistics regard this as having no grounds at all. The concept of predefined rank is considered an arbitrary notion, maybe useful for certain practical reasons but a source of problems when taken too seriously. As noted above, pruning and grafting often occurs, so that a class in the old situation sits under a genus in the new — such as happened with birds.

Modern phylogenetic nomenclature strictly and exclusively follows phylogeny and has arbitrarily deep and asymmetric trees.

More generally, there is no straightforward, objective measure of abstractness. So if you want to classify your assets in socks and other garments, or solar systems and galaxies, or Balkan destinations and Asian destinations, asymmetry is not an argument against doing so.

6 Conclusion

We started this paper by observing that the domains of search and business process management converge. The examples of website newtoholland.nl and the knowledge models supporting adaptive case management at the Dutch immigration service illustrate this: both use the same kind of knowledge models to reason based on business rules and the context of a specific user or case. Taxonomies form the basis of these models — different taxonomies can be combined into an ontology. Depending on the purposes of a taxonomy, it can be either right or wrong, or rather useful or not useful. Three fundamental rules were discussed that govern the well-formedness and quality of taxonomies. These rules are independent but may interact in intricate ways. Finally, we discussed some rules that are sometimes considered important but are mostly unjustified. In Part II and Part III of The Taxonomy Revolution, we will investigate in more detail how these taxonomies and models based on them can be applied in the domains of findability and of business process management.

About Taxonic

We are Taxonic, an international and independent IT consulting firm. Our focus? Digital transformation and Knowledge Management, which we apply within two areas: Linked Data and Pega (Dynamic Case Management).

As a knowledge partner, we share our expertise, provide tailored advice, and are hands-on in projects from analysis to implementation. Our experts work with the latest technologies. Within the team, we have a broad range of experience; from advising on technology selection and integrating solutions into the technical environment to connecting the system landscape and complete implementation projects.

We offer a full end-to-end service that allows us to guide digital transformations from design to implementation, ongoing development, and maintenance. Additionally, we are actively involved in developing our own software products. We create innovative software solutions based on Linked Data and semantic technology.

Working with interdisciplinary teams enables us to tackle a wide spectrum of complex projects. Taxonic encourages the world to adapt to new technologies, and this is our contribution to society.

About Jan Voskuil

With a background as a linguist, Jan understands the concept structure and the importance of Linked Data and SKOS like no other. He has, therefore, been a strong advocate for the development of Linked Data in the Netherlands driven by his content-driven motivation. Jan enjoys reading and has an interest in geology, which is not an unusual combination for an information architect.

Jan is a "thought leader" with several publications to his name on Semantic technology and Linked Data. He has also put SKOS on the map in the Netherlands by advocating for its inclusion in the 'comply or explain' list.

Veldzigt 2
3454 PW De Meern
The Netherlands

T +31 (0)88 829 66 00
E info@taxonic.com
W www.taxonic.com

bank NL45RABO0161959660
VAT NL851339803B01