E-Business Vocabularies as a Moving Target: Quantifying the Conceptual Dynamics in Domains

Martin Hepp

E-Business and Web Science Research Group, Bundeswehr University Munich, Germany
mhepp@computer.org

Abstract. Most practically relevant domains include some degree of conceptual dynamics, i.e. new conceptual elements arise, and some old ones become irrelevant. This type of change imposes a substantial challenge on building up-to-date domain vocabularies, and models in general. In this paper, we (1) provide a generic simulation model, based on a simple colored Petri-net, for investigating the interplay between conceptual dynamics and model coverage and (2) quantify the dynamics in three selected areas, i.e., for computer components, pharmaceuticals, and for methods, recipes, and procedures in the inorganic chemical technology sector. We can show that all three areas undergo a substantial conceptual dynamics and that this may lead to weak domain coverage in respective vocabularies. Based on these findings, we (3) discuss approaches of how the engineering lag of building domain vocabularies can be reduced.

1 Introduction

Materializing the promises of semantics-aware systems will require high-quality domain vocabularies, i.e., such that cover respective topic areas in sufficient detail so that tasks like e.g. search, content extraction, content reuse and integration, or services discovery can be supported in real-world scenarios. For many domains of interest, we are unfortunately still lacking up-to-date vocabularies with a sufficient granularity and expressivity. Of those domain vocabularies that are published on the Web, only some are actively maintained and thus reflect the current domain vocabulary. Many others are rather outdated prototypes of one-time snapshots of a domain.

Now, we can observe that most real-world domains include some degree of conceptual dynamics, i.e., that new elements arise as some old ones become irrelevant. In the products and services domain, for instance, manufacturers are continuously inventing new types of goods, and technical progress requires adding new attributes; in physics, scientists can discover new types of particles or relations among them; and in the geopolitical domain, new states form and political borders change. In some cases, one may argue whether such change is really a change in the ontology (e.g. the domain theory) or rather one of the broader knowledge base (i.e. , including data). We will discuss this question in more detail in section 2, but assume that at least a part of those changes requires actually a change in the ontology. In the area of products and services, it is for instance pretty obvious that technological advancement will require new concepts or new attributes for describing novel product models or product instances.

A. Gangemi and J. Euzenat (Eds.): EKAW 2008, LNCS 5268, pp. 388–403, 2008.

Fensel has stressed that ontologies are the glue between real-world semantics and formal semantics [1], i.e. that they are not just formal domain theories but models of the world that must also reflect reality as *perceived* by human actors. On a philosophical level, we may argue over whether all abstract concepts exist independently of time and human discovery, e.g. whether the categories for all potential products and services that may ever be invented already exist. Practically, however, we can assume that ontology engineers and domain experts can only add such new elements to the ontology once they are *known and actively used* in the respective domain. In effect, the dynamics of conceptual elements in reality or in the perception of reality is relevant when building a domain ontology for a particular area of interest.

In database research, the problems caused by domain evolution have been stressed e.g. by Ventrone and Heiler [2]. Similarly in ontology research, the challenges of change and dynamics have already been discussed by several researchers, e.g. by Noy and Klein [3], Heflin and Hendler [4], Fensel [1], Klein and Fensel [5], and Klein at al. [6]. In the field of methodologies for ontology engineering, e.g. the DILIGENT approach has put iterative maintenance and user/domain feedback to the center of building ontologies in order to deal better with change [7].

Changes in ontologies have been traced back by Noy and Klein [3] and Klein and Fensel [5] to three causes, i.e. (1) changes in the domain, (2) changes in the (shared) conceptualization, and (3) changes in the explicit specification. Most research has focused on how the change in and among evolving formal specifications can be managed, e.g. how we can maintain interoperability in a network of changing ontology specifications (in the sense of formalizations) so that instance data and ontology imports remain consistent or at least so that conflicts are minimized.

Unfortunately, the order of magnitude and impact of conceptual dynamics of domains as the origin of change has not yet received a lot of interest from researchers. This is in sharp contrast to the fact that such dynamics may be a significantly limiting factor when building and maintaining detailed domain vocabularies. We assume that the lack of interest is likely because conceptual dynamics is less obvious when dealing with upper-level concepts such as "physical matter," "agent," or "intangible." Since, historically, finding ontological truth at a high-level of abstraction has been an important guideline of building ontologies in Computer Science, we may falsely assume that creating lasting ontologies was a mere matter of proper conceptual modeling. That is, once we have discovered a proper model of a domain of discourse, the conceptualization and specification would be stable for ages. While we fully agree that cleanliness in conceptual modeling is important for creating stable vocabularies, there is evidence that this alone will not be sufficient for dealing with the dynamics faced in domain- and task-specific vocabularies.

Pinto and Martins [8] were one of the few who identified dynamism as a relevant dimension of ontology engineering projects. Also, some work has been done with regard to measuring the amount of change in domain *specifications*: Klein et al. [6] briefly reported the amount of change in the UNSPSC categorization schema, and in [9], we presented a comprehensive analysis of the amount of update operations in the four products and services classifications UNSPSC, eCl@ss, eOTD, and the RosettaNet Technical dictionary. In eCl@ss, for instance, there have been about 280 new and more than 1200 modified classes *per 30 days* (!) in versions 5.x [9]. However, it must be stressed that these two studies did not attempt to measure the *domain* dynamics but the dynamics in domain *specifications*. This is insofar relevant as the dynamics in specifications is likely much lower than that in the actual domain. After all, specifications incorporate only that subset of the overall change which has successfully passed all bottlenecks and obstacles of the updating process. We know from [9] that even large industry classifications with more than 25,000 concepts still lack a lot of relevant concepts, and this despite the fact that the degree of formality of these specifications is very limited.

1.1 Our Contribution

In this paper, we (1) provide a generic simulation model, based on a simple colored Petri-net, for investigating the conceptual dynamics and the resulting domain coverage of vocabularies. We then quantify the dynamics in three selected areas, i.e. (2) for computer components, (3) pharmaceuticals, and (4) for methods, recipes, and procedures in the inorganic chemical technology sector. We can show that all three areas have substantial domain dynamics and that vocabularies updated in typical intervals will suffer from weak domain coverage. Consequently, we (5) discuss approaches of how building vocabularies in dynamic domains can be improved.

1.2 Structure of the Paper

In section 2, we analyze the interplay between domain dynamics, the vocabulary engineering and maintenance lag, and possible domain coverage of a vocabulary. In section 3, we present our research methodology, namely a simulation model based on a colored Petri-net. In section 4, we describe the data sources that we used for our experiments and what kind of pre-processing we carried out. In section 5, we present the results from the simulation runs and show how the amount of concepts missing in a respective domain vocabulary would develop over time. In section 6, we discuss and evaluate our findings and derive implications for building domain vocabularies. Section 7 highlights our main results and concludes the paper.

2 Conceptual Dynamics and Vocabulary Maintenance Lag

In this section, we analyze how the unavoidable delay in producing a shared formalization constrains the inclusion of novel conceptual elements from the domain, and thus limits the amount of up-to-dateness of a domain vocabulary.

2.1 Conceptual Elements in Domain Vocabularies

Vocabularies are commonly specified as a set of conceptual entities of the respective domain of discourse. If at the level of ontologies, the vocabulary elements are usually (1) described using human readable text and (2) their interpretation is constrained by formal axioms [cf. 11, 12]. The typical elements of ontologies are classes, attributes (slots), relations, functions, and axioms. Also, instances (individuals) may belong to the ontology as long as they are “ontological” instances, i.e. such that are not mere data but a necessary part of the shared conceptualization. It highly depends on the scope and purpose of the ontology whether a particular individual is an “ontological instance" or data of the knowledge base, and there is often also room for argument. However, there is little doubt that some instances belong into the ontology. This is in particular true for *domain* ontologies and is insofar relevant as much of the domain dynamics takes place in the area of "ontological instances". For the remainder of the paper, we refer to all potential elements of an ontology as "conceptual elements" and mean with this all concepts/classes, attributes, relations, *ontological* instances/individuals, functions, and axioms.

2.2 Model of Formalization Delay and Domain Coverage of Vocabularies

Whenever we design an domain vocabulary, we face a fundamental problem: It takes time for the involved stakeholders (1) to agree upon the relevant conceptual elements and their definition and (2) to produce a shareable formalization. At the same time, new conceptual elements become relevant in a domain of discourse, which was not yet included in the initial domain capture. From an engineering perspective, it would be better if we were able to „freeze“ the discourse and dynamics while we are working on the consensual model of the domain, but of course we can not. This holds both for the initial formalization of an ontology and for consequent updates. The very same problems are known from standardization [13]. They have also recently been sketched for ontologies [14].

We can find such dynamics in almost any domain: in a sports and leisure ontology, new types of sports activities are becoming popular (e.g. "rafting", "sandboarding", "street skating", "kite snowboarding", etc.). In the legal domain, new categories of punishable acts may be defined. In biology and medicine, new classes of species may develop due to evolution or may be discovered and named. Now, it is a triviality that a domain's conceptual dynamics increases with the specificity of modeling, (a class hierarchy's granularity, for example). As long as (1) the domain model is fairly abstract, (2) the engineering and maintenance delays are small, and (3) the conceptual dynamics is limited, updating the specification is a lesser problem. Since ontologies should follow the principle of minimal ontological commitment [15] and thus be, in general, more abstract, this type of a knowledge acquisition and maintenance bottleneck is less problematic than with detailed knowledge bases. Eventually, the conceptual dynamics will be very limited when building top-level ontologies, since the microscopic everyday advancements that mankind makes usually leave the big categories of "tangible vs. intangible" or "role vs. actor" untouched.

However, it is also pretty obvious that materializing most of the promises about ontologies and the Semantic Web will require *detailed* domain vocabularies in addition to top-level abstractions. For instance, if we want to use ontologies for the automated mediation between message flows from a set of incompatible legacy systems in the billing processes of telecommunications companies, then we need such ontologies that reflect all the conceptual elements in that domain: contract types and billing plans, locations, types of telecommunication equipment, etc. Likely the most striking example is the products and services domain, where new types of goods and new attributes of existing goods categories are invented or introduced on a daily basis.

One could argue that it was just a matter of fact that some novel concepts were missing in a domain vocabulary. Unfortunately, often the novel concepts in a domain are those for which semantic technology would be most interesting. For instance, when searching the Web for price comparison, it is the novel goods for which the price differences are likely most substantial. For long established categories, competition on the market and arbitrage will have balanced out prices, and traditional technology like search based on controlled lexical resources may be used to facilitate search. In comparison, ontologies could help tremendously when broad consensus about terminology has not yet been established.

The basic structure of the problem of domain dynamics on one hand and the lag in vocabulary creation and maintenance on the other hand is illustrated in Figure 1. The upper line depicts the amount of conceptual elements in reality, caused by the continuous "birth" of new ones. (For the moment, we abstract from the removal of outdated elements, since keeping them in the vocabulary does often not harm.) The lower line reflects the amount of conceptual elements that are included in the most recent release of the vocabulary. We can easily see the fundamental problem: Once the initial domain capture for the ontology is completed (t_0), it takes some time to formalize and release the ontology. Thus, the first version of the ontology will not be available until t_1. This vocabulary contains all elements (classes, instances, attributes, relations, and axioms) that reflect the initial domain capture in t_0. In the meantime, however, additional conceptual elements have become relevant in the real world, as depicted by the upper line. All such new elements are not included in the vocabulary and can thus neither be used for annotating data nor for expressing queries.

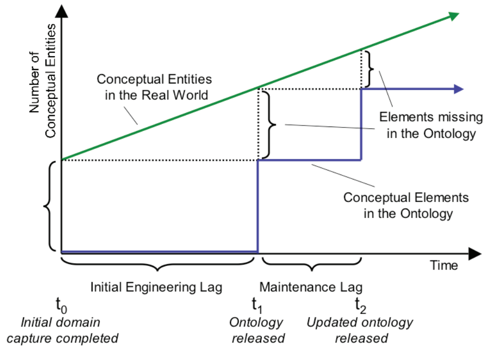
Number of Conceptual Entities
Conceptual Entities in the Real World
Elements missing in the Ontology
Conceptual Elements in the Ontology
Time
Initial Engineering Lag
Maintenance Lag
t0
Initial domain capture completed
t1
Ontology released
t2
Updated ontology released

Fig. 1. Conceptual dynamics and coverage of domain vocabularies

If the ontology is actively maintained, we may carry out an updated domain capture at t_1, but producing the updated vocabulary and documentation again takes time, making the new version available at t_2. In the meantime, however, additional conceptual elements have again become relevant in the real world, which will be missing in this vocabulary update. Thus, we are trapped in a vicious circle: Each time we release a new version of the vocabulary, it may be the perfect shared conceptualization for the respective domain – but only with regard to the state in which the domain was when we completed our domain capture.

3 Methodology

In this section, we describe our research methodology and the simulation model.

3.1 Overview

In order to quantify the impact of conceptual dynamics in a domain on the coverage of current elements in a vocabulary, we designed a simulation model, based on a colored Petri-net. For an overview of Petri-nets and their application to simulation, see e.g. [10]. Petri-nets consist of places, transitions, directed arcs, and tokens. Dynamic behavior in Petri-nets is basically represented in a Petri-net by the consumption of input tokens and the creation of new output tokens by a transition (this process is known as "firing" of a transition). That means that the structure of the net is defined at build-time of the simulation and the dynamics is represented by a flow of tokens through the net at run-time.

Simple Petri-nets allow only one token per place; a transition will fire if all input places contain exactly one token and all output places are empty. In this case, one token from each input place will be consumed and one on each output place will be created.

There are numerous extensions of Petri-nets, in particular "colored" Petri-nets and such that model time explicitly. In colored Petri-nets, tokens can be distinguished in that each token may carry certain properties. The transitions may contain conditions that refer to the properties of tokens on the input places. When a transition fires and consumes tokens from input and creates tokens in the output, the transition may assign any kind of locally generated results to the created tokens on the output places.

For Petri-nets, there is a wealth of simulation environments available. For our experiments, we chose the commercial package PACE[1] version 4.0. PACE is based on colored Petri-nets and uses Smalltalk as the language for specifying conditions and other instructions inside the transitions.

3.2 Simulation Model

The basic idea of our model is as follows:

1. From domain data, we extract a set of past events that can be regarded as the birth of a new conceptual element, and store the day of its first appearance.
2. If possible and reasonable, we obtained or estimated the lifespan of the conceptual entity, i.e. the duration for which the element would belong to the active vocabulary. Though we do not need to remove non-conflicting, outdated elements from a domain vocabulary, such may be helpful because we can use it later for determining the domain coverage of a vocabulary as a percentage over the elements in the current state of the real world – e.g. how many concepts are included in the vocabulary vs. how many are used in the real world.
3. We assume that the time behavior of the vocabulary engineering and maintenance process can be approximated using one of the following patterns or a combination of those:

[1] http://www.ibepace.com/

a) **Regular update in fixed time intervals:** A new vocabulary version is produced at regular points in time. This new version will include all conceptual elements that are waiting in the processing queue. In order to make the model more realistic, we also introduced lead time, i.e., elements must arrive a certain amount of days before the next release in order to be included. Those that arrive later will only be processed in the update following the upcoming one.

b) **Fixed processing capacity per time:** A new vocabulary version is produced as in a), but additionally, the processing "bandwidth" for the vocabulary is limited in that only a maximum number of change requests can be processed within a given amount of time.

c) **Minimum amount of change requests:** A pre-defined minimum of change request must be waiting in the queue for triggering the update of the vocabulary.

Since the process of adding the elements is modeled as one transition, it can be easily extended so that it reflects the time behavior in more detail. For example, multiple stages or voting and review mechanisms can be added easily.

Following that guideline, we designed the simulation model as shown in Figure 2. The place P1 will hold all tokens that represent such conceptual elements that will become relevant during the analyzed time-span. For example, we created tokens for patent applications that reflect novel procedures or materials. Transition T1 will fire once the date of birth of a token in P1 is reached. It creates two new tokens, one on P3 and one on P2. The place P3 reflects the set of conceptual elements in the real world – quite trivially, these are all tokens from P1 that have already been "born", minus those that have already become irrelevant. The latter process is represented by T5: Once the current time is greater or equal to the date of birth plus the lifespan of a conceptual element, it is removed from P3.

The place P2 reflects the maintenance queue of the vocabulary engineering process, i.e., such elements that have been added to the real world but are yet to be incorporated in the next vocabulary release. The process of adding a waiting element to the next vocabulary release is represented by T2. The conditions for this transition can be set to reflect the various update patterns a), b), or c) as described above. Once T2 fires, it consumes the waiting conceptual element from P2 and creates a new element on P4. The place P4 represents the current vocabulary release, i.e., the set of conceptual elements that have yet been added to the vocabulary. Same as in the real world, outdated conceptual elements are removed from the vocabulary via firing of T3 if the lifespan has lapsed. We did not model explicit removal and a new lag in here, since the removal of outdated conceptual elements is used only to be able to determine the domain coverage of the vocabulary as a percentage of the current domain elements.

In the model, we represent time by discrete time tokens waiting in P5. Each action that takes place at a given moment in time consumes such a time token and creates and returns a new one immediately. This is depicted by the double-lined arcs with arrows at both ends between P5 and the respective transitions (this is a common notion in several Petri-net tools). Only if no other transition (e.g. T1, T2, T3, or T5) consumes the time token waiting on P5, transition T4 will fire and create two new tokens, one new time token with an increased time value "clock" on P5 and one

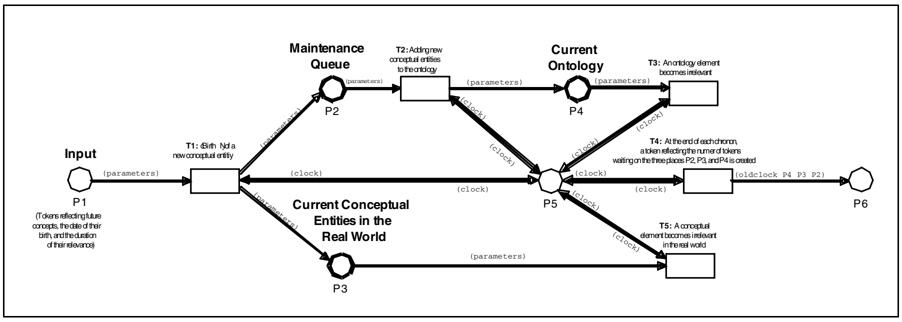
Input
P1
(Tokens reflecting future concepts, the date of their birth, and the duration of their relevance)
(parameters)
T1: dBirth Ņof a new conceptual entitiy
(parameters)
(parameters)
Maintenance Queue
P2
(parameters)
T2: Adding new conceptual entities to the ontology
(parameters)
Current Ontology
P4
(parameters)
T3: An ontology element becomes irrelevant
(clock)
(clock)
(clock)
(clock)
(clock)
(clock)
T4: At the end of each chronon, a token reflecting the numer of tokens waiting on the three places P2, P3, and P4 is created
(clock)
(clock)
(oldclock P4 P3 P2)
P5
P6
Current Conceptual Entities in the Real World
P3
(parameters)
(clock)
(clock)
T5: A conceptual element becomes irrelevant in the real world

Fig. 2. Petri-net model of conceptual dynamics and vocabulary maintenance

token on P6. The priority of all other transitions is implemented by a delay in the definition of T4. The newly created token on P6 has as attributes the current amount of tokens waiting on P2, P3, and P4 plus the respective time. These tokens reflect the amount of conceptual elements available in the real world (P3), those already contained in the current vocabulary (P4) and those waiting in the vocabulary maintenance queue (P2) for every discrete moment in time within the simulated time-span. They can be used easily to draw time series diagrams for those values.

The "clock" property of the time tokens is stored as a positive integer value indicating the number of days from January 1, 1900 on (i.e. 1 = Jan 1, 1900; 2= Jan 2, 1900, …).

3.3 Model Calibration

For all further experiments, we assumed that a new vocabulary release is produced every 360 days and that it will include all waiting change requests as long as they are available at least seven days before the update. Of course, shorter update cycles may be possible in some vocabulary projects. A quick look on ontologies on the Web showed however that one comprehensive update per year is rather the notable exception than the rule. Accordingly, we set the condition for T2 to:

```
(clock\\360=0) & (clock >=((parameters at: 2) + 7))
```

4 Data Sources and Pre-processing

Our simulation model expects input data as a set of triples of the form `(ID, dateOfBirth, lifeSpanInDays)`. That means that the ideal type of input data would be a log file of "birth" dates of all conceptual elements in a clearly defined domain of discourse. Each entry should specify the date on which the element was introduced to the domain and indicate the time-span for which this element belongs to the active vocabulary. Also, the log file should cover at least multiple times the average lifespan of a single conceptual element.

Since nobody keeps such a complete register of the birth and death of conceptual elements, respective data is not readily available. Thus, we derived the input data for our experiments from reasonable proxies for the appearance of new conceptual elements. There are at least three promising types of sources for such data:

1. Public announcements of relevant concepts or ontological instances, e.g. the official release data of the first product of a new kind.
2. Patent applications, in which an entity seeks protection of the exploitation rights for a new device, method, or material.
3. Filing data or public notices of applications for public approval, e.g. such in which an entity seeks permission for a new type of treatment or the introduction of a new type of medical product.

Alternatively, one could generate random input data based on assumptions about the distribution properties. However, since such distribution properties are hard to determine and hard to justify, and since we were able to find promising real data sources, we did not consider this option.

For our simulation, we tried to find data sources for individual areas that are coherent enough to be considered a domain of discourse and for which the creation of a domain vocabulary would very likely provide business benefit. We were successful in obtaining such data for three selected areas as following.

4.1 Intel CPUs

Intel CPUs are important conceptual elements when describing the market for computers and computer peripherals. This is because each individual Intel CPU make and model (e.g. “Pentium III 300 MHz”) is helpful for describing the performance and characteristics of a desktop or laptop computer. For makers of motherboards or other components, being able to specify the supported types of CPUs by referring to a vocabulary will also provide benefits. Thus, it is pretty obvious that Intel CPU types are relevant conceptual elements of the computer parts domain. Also, a particular Intel CPU is more specific than just a CPU for which the manufacturer is “Intel” and the clock speed is e.g. 2 GHz – each CPU model is a conceptual entity in its own right.

From [16], we were able to obtain a complete list of the release dates of all Intel CPU models back to the famous model 4004 released in 1971. Table 1 shows an excerpt from the respective data. For our experiments, we focus on the conceptual dynamics between January 1, 1997 and January 1, 2002. As a speculative extension of the experiment, we estimated the “conceptual lifespan” of such CPU models, i.e., for how long they actually belong to the active vocabulary. Due to the high degree of innovation in markets, outdated CPU models are for example almost completely irrelevant when describing offers in the E-Commerce domain. We guessed that CPUs released until the end of 1997 would belong to the relevant concepts for 720 days after their introduction and those introduced from 1998 onwards for 360 days. These estimates are based on a subjective assessment of the pace of innovation on the PC market, and on the increased amount of new releases after 1998. Of course one can argue that there will always be some old data referring to past CPU models and thus such estimates are always questionable. However, we can reasonably assume the biggest share of offerings data in the e-commerce domain to refer only to *current* CPUs. Also, the proposed extension is not necessary for the main experiment. It will, however, allow us to measure the domain coverage as a *percentage* of vocabulary

Table 1. Sample from the Intel CPU data (based on [16])

Release Date (“Date of Birth”)	Release Date (Integer Value)	Intel CPU Model
26.01.1998	35821	Pentium(r) II Processor (333 MHz)
02.04.1998	35887	Mobile Pentium(r) II Processor (233 and 266 MHz)
15.04.1998	35900	Intel(r) Celeron(r) Processor (266 MHz)
15.04.1998	35900	Pentium(r) II Processor (350 and 400 MHz)
08.06.1998	35954	Intel(r) Celeron(r) Processor (300 MHz)
29.06.1998	35975	Pentium(r) II Xeon(tm) Processor (400 MHz)

elements over current conceptual elements in the real world (i.e.: “how much world is in the vocabulary”).

4.2 Pharmaceuticals

In the field of pharmaceuticals, it was rather difficult to obtain meaningful data. Eventually, we decided to take the filing data of applications for the approval of new pharmaceuticals in the United States as our main data source. As raw data, we used the drug application data from the US Food and Drug Administration (FDA) from July 1996 through July 2002, which we could obtain from [17] in the form of HTML files. The FDA applications are divided in categories as follows [cf. 17]:

a) “Original New Drug Applications”,
b) “Efficacy Supplemental New Drug Applications”,
c) “Approvable Original New Drug Applications”,
d) “Original Abbreviated New Drug Applications”,
e) “Original Abbreviated and 505(b)(2) New Drug Applications with Tentative Approval”, and
f) “Labeling Supplements to Original New Drug Applications”.

For our analysis, we only considered category a), since only these are truly novel drugs. Applications in this category are described in more detail using the attribute “chemical type”. We removed all entries that are of the subtype „Already marketed drug but a new manufacturer“ (type 5, see [17], 20 entries) und two obvious redundancies (#21015, ANDROGEL) and (#21124, LAMISIL). All in all, this returns 481 valid elements.

Table 2. Example of the FDA pharmaceutical data (based on [17])

Original Application No.	Approval Date	Tradename
20616	03. Jul 96	KADIAN
20536	03. Jul 96	NICOTROL
20630	12. Jul 96	ULTIVA
50711	18. Jul 96	ZITHROMAX
20554	22. Jul 96	DOVONEX
20625	25. Jul 96	ALLEGRA

This selection of data corresponds to a vocabulary of pharmaceutical substances and products. This could for example be used to support procurement processes of pharmaceuticals, annotate prescription data for mining purposes, or semantics-supported healthcare applications, etc. We retrieved all monthly reports, merged them into one large HTML file, and extracted the plain text data. Then, we used a small Java program to export the relevant data fields and write them to a CSV file. Table 2 shows a sample from the respective data.

4.3 Methods, Materials, and Procedures in Inorganic Chemistry

As the third segment in our analysis, we looked at innovation in the inorganic branch of chemistry. The starting point for our data were German patent applications. We assume that patent applications are a good estimate for the lower limit of the conceptual dynamics in this domain, since the application for a patent is associated with cost and effort. It is thus safe to assume that the applicants expect business significance of the innovation – both economically and conceptually.

Since collecting the data from a Web database showed to be very labor-intensive, we constrained our analysis to the branch of *inorganic* chemistry. The source of our data was the database of the German patent and trademark registry [18]. We retrieved pending and approved applications for patents and utility models and regarded each application as a surrogate for a novel conceptual element in the domain. The system supports search by patent categories using International Patent Classification (IPC) codes. The IPC is a hierarchical classification schema for patents. We selected section C „Chemistry; Metallurgy", subsection C01 („Inorganic Chemistry") [19]). Within that subsection, we considered the segments C01B, C01C and C01D. Those three are defined as given in Table 3 ([cf. 19]):

Table 3. Definition of the IPC patent categories C 01 B, C, and D

IPC	Definition
C 01 B	Non-metallic elements; compounds thereof
C 01 C	Ammonia; cyanogen; compounds thereof
C 01 D	Compounds of alkali metals, i.e. lithium, sodium, potassium, caesium, or francium

We queried the database for the period from January 4, 1999 – December 15, 2002 on a weekly basis. In order to keep the amount of queries in a reasonable order of magnitude, we used the last day of the respective week as the „birth date" and not the actual date of the application. However, we assume that this small deviation can be neglected. All in all, we carried out 624 manual queries (52 weeks * 4 years * 3 patent categories).

We took into account all entries from the section „patent applications and utility models", using the date of the application document as the date of birth for the respective conceptual element. All in all, we collected 490 application documents for patents and utility models for the given period of time. In this case, we manually extracted all data from the HTML files per week. We then wrote a small Smalltalk script for creating respective tokens on place P1 in the simulation environment. The simulation run started with January 10, 1999 (birth of the first element, day 36170 as an integer) and ended on December 15, 2002 (day 37605 as an integer).

5 Results

In this section, we summarize the results from the three simulation runs and discuss on how the conceptual dynamics in the selected domains influences the construction of domain vocabularies. As said, we assumed one vocabulary update per each 360 days with a deadline for inclusion of seven days before the respective update takes place, which is the lower limit of release delays in many standardization processes. Table 4 gives an overview of the amount of missing elements in a respective domain vocabulary created in such a setting.

Table 4. Amount of missing elements in the three domains

Domain	Time-span analysed	Amount of Missing Elements		
		Mean	Median	Max
Intel CPUs	1/1997 – 1/2002	6.835	5	24
Pharmaceuticals	7/1996 – 6/2002	43.689	38	129
Inorganic Chemical Innovations	1/1999 – 12/2002	49.321	45	150

We can see from the median value that during half of the time, at least five Intel CPU concepts, 38 pharmaceutical innovations, and 45 concepts reflecting new methods, materials, or procedures in the inorganic chemical industry sector, are missing in the respective vocabulary. Shortly before the next vocabulary update, the number of missing elements rises up to 24, 129, and 150 respectively. It is important to stress that those missing elements reflect the innovative part of the domain, which has usually much higher business relevance for search and information processing tasks. Not being able to use semantic technology for processing data that refers to those "hot topics" may drastically reduce the business value of semantic technology in the respective domains. For example, few people searching for a place that sells bread or butter will consult the Web, as compared to someone searching for "wakeboarding on Mauritius" or a "Bluetooth noise-canceling headset for Nokia". Similarly in the B2B segment, a producer of pharmaceuticals may want to use Semantic Web technology for watching all news and blog entries referring to a novel type of product of the competition, and a semantics-supported knowledge base for customer support of PC manufacturers may require that we annotate incidents using an vocabulary of the involved CPU model. In the following, we present the detailed simulation results per each domain.

5.1 Intel CPU Vocabulary

As for an Intel CPU vocabulary, we can see clearly that the conceptual dynamics in this domain is increasing year by year. During at least half of the time, five or more such concepts are missing in the domain vocabulary (median = 5); shortly before the annual update, this rises to a maximum of 24 missing entries.

In this domain and based on the assumptions described in section 4.1, we were also able to estimate the *full* size of the active vocabulary in such a vocabulary. If we use these lifespan estimates, then we can determine the median number of CPU model concepts in the active vocabulary as 19; the mean is 16.961. This says that about half of the time, five or more of the CPU types in an active vocabulary (of on average) 17 CPU types are not yet included in the vocabulary. And, unfortunately, those are likely the most interesting ones for business entities, and those for which Semantic Web-based comparison-shopping would be most attractive due to high price dispersion among novel products.

We also computed the domain coverage for each day in the simulation run. For this, we divided the amount of elements in the domain vocabulary (i.e. the number of tokens on place P4) by the amount of elements in the real world (place P3). For the series of the resulting values, the mean is 61.55 %. In other words, on average almost 40 percent of current CPUs would be missing in the vocabulary.

5.2 Vocabulary of Pharmaceuticals

As for a vocabulary of FDA-approved pharmaceuticals (simulation period July 3, 1996 through June 26, 2002), we see a slight decrease in the conceptual dynamics year by year. Also, when compared to the overall amount of FDA-approved drugs, the average of 43 missing elements may be acceptable. This reflects that, due multi-stage clinical trials and in general the long time-to-market of new pharmaceuticals, the dynamics in this domain is heavily constrained by legal regulations.

5.3 Vocabulary of Methods, Materials, and Procedures in Inorganic Chemistry

As for a vocabulary of methods, materials, and procedures in inorganic chemistry, the conceptual dynamics in this domain is increasing year by year. This may reflect either a general increase in research and development productivity, or just an increase in the tendency to seek patent protection. During at least half of the time, 45 or more such concepts are missing in the domain vocabulary; shortly before the annual update, this rises to a maximum of 150 missing entries.

The total number of innovations in the domain of discourse over the full duration of the experiment was 490. For reasons of comparison, we also determined how the amount of missing elements would be reduced if the vocabulary was updated every 21 days (still requiring seven days lead time). In that case, the mean would be just 6.74 missing elements. Since we can assume that the majority of innovations in this domain will have a lifespan that by far exceeds the duration of our simulation run, we cannot determine the absolute size of the current vocabulary in the real world.

6 Related Work

Steels and Kaplan carried out simulation experiments on how a group of autonomous agents can update their shared vocabulary so that it incorporates semiotic dynamics in the community of these agents [20]. Fensel stressed that building ontologies solely based on the search of ontological truth is problematic, because human actors are able to find shared domain conceptualizations only in a social process by means of perception and argument, in which the conceptualization is both means and object [1]. Thus, Fensel claims that dynamics in the domain is caused not only by factual, objective changes but also by progress in the argument and changes in the perception of the world. Currently, there is a lot of interest in re-using the Wiki-approach as an ontology engineering environment. This has in common with our work the assumption of domain dynamics as a focal point of ontology building. Our findings give additional evidence for the justification of this assumption.

Oliver et al. analyzed how change in medical terminology caused by scientific advancement can be managed in controlled vocabularies [21]. Schulze and Stauffer carried out simulation experiments on the diffusion of languages based on individual variation, passing along, and selection [22]. In contrast to our work, their analysis focuses on the dynamics of the adoption of languages as larger units, while we address dynamics at the level of individual concepts. Additionally, earlier experiments of ours have been reported in [13]. We do not know of any other quantitative research on the interplay of domain dynamics and vocabulary coverage.

7 Discussion and Conclusions

We provided a generic Petri-net-based model for relating the dynamics in a domain to producing and maintaining conceptual models for that domain. One of the properties of the Petri-net approach is that the temporal behavior of the vocabulary maintenance process can be modified or refined as needed, so that it would reflect various modes of maintenance, e.g. fix intervals, bandwidth/capacity constraints, or delays in a multi-stage process with explicit voting etc. Also, the simulation model itself can be applied to any conceptual modeling problem.

The simulation runs clearly show that there is a substantial amount of conceptual dynamics in the three domains. In the case of an Intel CPU vocabulary, the results are most obvious. First, the benefits of a respective domain vocabulary are easy to identify. Second, based on assumptions about the duration of domain relevance of a new processor model, we could even determine the average domain coverage in percent. All in all we can see that the order of magnitude of domain dynamics poses a challenge for building current domain vocabularies, and that this challenge is on top of the technical challenges addressed by available infrastructure for ontology versioning and evolution. Unfortunately, we can assume that in many application domains, the most novel concepts must be available in the vocabulary in order to exploit the business potential of semantic technology.

In dynamic domains, the possible degree of domain coverage is constrained in two ways: First of all, the group of individuals building the vocabulary must be aware of the novel conceptual elements. In here the bottleneck is often whether users of the vocabulary have an easy-to-use mechanism of reporting missing elements. Secondly, the lag and quality of the vocabulary maintenance process limits the inclusion of such change requests. This is a strong indicator that in domain vocabularies, the problems of vocabularies constructed by a small “elite” but meant for a bigger set of users is more problematic than for abstract upper-level ontologies [cf. 14].

Of course, we can mitigate most of the problems by increasing the level of abstraction. However, materializing most of the promises of the Semantic Web will require very detailed domain ontologies. After all, most of the data exchanged in commerce is referring to very specific categories of things – few people order „hot beverages“ in a bar (rather „a café latte with macadamia flavor“) and employment agencies do not search for „human actors“ (rather „mechanical engineer with >= 3 years of professional experience in maintaining car-wash systems“).

The only feasible approach for dealing with dynamic domains is speeding up vocabulary maintenance and taking away obstacles for grasping user feedback for extensions and corrections. It is obvious that monthly or weekly updates of the vocabularies in our simulation experiments will drastically reduce the amount of missing elements. However, the bigger the number of stakeholders involved in the conceptualization and formalization of a vocabulary, the more time will be necessary for reaching agreement. In the following, we summarize findings from our experiments:

(1) Not only the absolute duration of the update interval is relevant, but also the relative timing of the release dates. This is in particular true if the usage of the vocabulary will be unevenly distributed over time, e.g. due to seasonal effects.
(2) Proper conceptual modeling alone does not solve the problem of domain dynamics.
(3) The group of individuals taking care of the maintenance of a vocabulary must establish mechanisms that make it as simple as possible and as rewarding as possible for plain users to report change requests. Otherwise, missing elements may be spotted but never reported.
(4) We need to think of vocabulary modularization also in terms of decoupling domain dynamics and distributing responsibility. In this respect, ontology engineering can learn from the lessons in creating lasting numbering schemas like EAN/UPC or the ISBN, for which the standardization bodies assign authority over the subsets of the naming space according to a hierarchical schema. This allows e.g. any company in the world to define globally unique EAN/UPC codes within their branch without delay.
(5) We may be able to predict the emergence of a novel concept. In that case, already the early indicators should trigger the change request.
(6) In some cases, lightweight ontologies that capture evolving classification schemas as pure literal values may be a better solution than to continuously replicate the schema as an ontology formalization.

Insufficient domain coverage is a major problem since it will often be the very new conceptual elements in domain vocabularies that empower semantic systems to provide business benefit in terms of agility and operational efficiency. Current approaches of bringing domain ontology engineering back into the hands of the users, e.g. on the basis of Wiki technology, are likely an important direction.

Acknowledgments. Part of the work presented in this paper has been supported by the Austrian BMVIT/FFG under the FIT-IT Semantic Systems project myOntology (grant no. 812515/9284).

References

1. Fensel, D.: Ontologies: Dynamic networks of formally represented meaning (retrieved December 15, 2007), http://sw-portal.deri.at/papers/publications/network.pdf
2. Ventrone, V., Heiler, S.: Semantic heterogeneity as a result of domain evolution. ACM SIGMOD Record 20(4), 16–20 (1991)

3. Noy, N.F., Klein, M.: Ontology Evolution: Not the Same as Schema Evolution. Knowledge and Information Systems 6(4), 428–440 (2004)
4. Heflin, J., Hendler, J.: Dynamic Ontologies on the Web. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI), Austin, Texas, pp. 443–449 (2000)
5. Klein, M., Fensel, D.: Ontology versioning for the Semantic Web. In: Proceedings of the International Semantic Web Working Symposium (SWWS), Stanford University, California, USA, pp. 75–91 (2001)
6. Klein, M., Ding, Y., Fensel, D., Omelayenko, B.: Ontology Management: Storing, Aligning and Maintaining Ontologies. In: Davies, J., Fensel, D., Harmelen, F.v. (eds.) Towards the Semantic Web, pp. 47–69. Wiley, Chichester (2003)
7. Vrandecic, D., Pinto, S., Tempich, C., Sure, Y.: The DILIGENT knowledge process. Journal of Knowledge Management 9(5), 85–96 (2005)
8. Pinto, H.S., Martins, J.P.: Ontologies: How can They be Built? Knowledge and Information Systems 6(4), 441–464 (2004)
9. Hepp, M., Leukel, J., Schmitz, V.: A Quantitative Analysis of Product Categorization Standards: Content, Coverage, and Maintenance of eCl@ss, UNSPSC, eOTD, and the RosettaNet Technical Dictionary. Knowledge and Information Systems 13(1), 77–114 (2007)
10. Peterson, J.L.: Petri Nets. ACM Computing Surveys 9(3), 223–252 (1977)
11. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2), 199–220 (1993)
12. Noy, N.F., Hafner, C.D.: The State of the Art in Ontology Design. AI Magazine 18(3), 53–74 (1997)
13. Hepp, M.: Güterklassifikation als semantisches Standardisierungsproblem. Deutscher Universitäts-Verlag, Wiesbaden (2003)
14. Hepp, M.: Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies. IEEE Internet Computing 11(7), 96–102 (2007)
15. Uschold, M., Grüninger, M.: Ontologies: Principles, Methods, and Applications. Knowledge Engineering Review 11(2), 93–155 (1996)
16. Intel Corp.: Intel Microprocessor Quick Reference Guide (retrieved December 15, 2006), http://www.intel.com/pressroom/kits/quickrefyr.htm
17. U.S. Food and Drug Administration: FDA Drug Approvals List (retrieved December 18, 2006), http://www.fda.gov/cder/da/da.htm
18. Deutsches Patent- und Markenamt: DPINFO. Datenbank zu Patenten, Gebrauchsmustern, Marken und Geschmacksmustern (retrieved December 15, 2006), https://dpinfo.dpma.de/index.html
19. World Intellectual Property Organization: IPC 7 English Version Section C (retrieved January 10, 2003), http://www.wipo.org/classifications/fulltext/new_ipc/ipc7/ec.htm
20. Steels, L., Kaplan, F.: Collective learning and semiotic dynamics. In: Floreano, D., Mondada, F. (eds.) ECAL 1999. LNCS, vol. 1674, pp. 679–688. Springer, Heidelberg (1999)
21. Oliver, D.E., Shahar, Y., Shortliffe, E.H., Musen, M.: Representation of change in controlled medical terminologies. Artificial Intelligence in Medicine 15(1), 53–76 (1999)
22. Schulze, C., Stauffer, D.: Monte Carlo simulation of the rise and the fall of languages. International Journal of Modern Physics C 16(5), 781–787 (2005)