Measuring the Quality of Descriptive Languages for Products and Services

Working Paper 1/2004

Dr. Martin Hepp
Department of Computer Information Systems
College of Business
Florida Gulf Coast University
10510 FGCU Boulevard South
Fort Myers, FL 33965-6565, USA

Phone: (+1) (239) 590-7311
Fax: (+1) (239) 590-7330

E-Mail: mhepp@computer.org

Preferred Citation:
Hepp, Martin: Measuring the Quality of Descriptive Languages for Products and Services. Working Paper 1/2004, available at http://www.heppnetz.de/metrics-hepp.pdf.

Measuring the Quality of Descriptive Languages for Products and Services

Martin Hepp
Florida Gulf Coast University, Department of Computer Information Systems
mhepp@computer.org

Abstract: Standardized product classification systems play a major role for automated business transactions. They are not only a coding scheme that eases catalog data integration, but can act as descriptive languages for products and services, supporting a multiplicity of future e-business scenarios. Even though prominent approaches like eCl@ss and eOTD are built around a hierarchical order of product and service categories, this is not a compulsory property. The key functionality is the representation of business meanings (e.g. a product or service) in an unambiguous, machine processable manner. Currently, the most popular architecture for such descriptive languages is the combination of three components: classes, an attribute library, and class-specific attribute lists, which consist of references to characteristic attributes for the respective class. The development and maintenance of the respective attribute lists is a major challenge, as this requires comprehensive domain knowledge about the respective goods. Eventually, the quality and usefulness of the descriptive language is determined by the quality of its attribute lists. This paper proposes metrics that help measure the quality of and the progress in the development of descriptive languages for products and services.

Keywords: Product classification, eCl@ss, eOTD, UNSPSC, attribute lists

Contents

1 Introduction

Standardized product classification systems like eCl@ss play a major role in business related communications between systems, because they allow the unambiguous and context-free encoding of products and services. The most established area of application is catalog data exchange and consequent catalog data integration, where product data from multiple sources must be assigned to categories of a target catalog structure (cf. Baron et al. 2000; Agrawal / Srikant 2001; Fensel et al. 2001; Stonebraker / Hellerstein 2001; Leukel et al. 2002). A common form of this task is grouping product data sets from multiple vendors into vendor independent categories.

Though many prominent approaches like eCl@ss (www.eclass.de), UNSPSC (www.unspsc.org), and eOTD (EGIS/EGAS, see www.eotd.org) are built around a hierarchy of product and service categories, this is not a compulsory property. The key functionality is the representation of business meanings (e.g. a product or service) in an unambiguous, machine processable manner. This need not reach as far as a true ontology (cf. Guarino 1998; Fensel 2001), which usually contains a formal definition of each concept's properties (including its relationships to other concepts in the ontology). For many business applications, it is sufficient if the concepts behind the entries are defined in a semi-formal way or even in natural language. In other words, the semantic richness of the entries is not as important as the semantic precision. The reason is that most practically relevant business applications will require only very basic inference operations, at least in the foreseeable future. For example, it is important for a machine to determine whether product 1 is a substitute for product 2 or whether it belongs to product category A, but it is not yet needed to draw more complex conclusions.

The commonly used term "product classification system" fosters a false perception of the nature and future relevance of this field of research. First of all, the term "classification" is frequently used with "hierarchical classification" in mind, despite the fact that classification itself does not require a hierarchy. Second, the key task is not to divide a set of products into subsets, but to provide symbols (signs) that can be used for labeling objects.

On the other hand, it does not make sense to stress the ontological aspect by using a term like "product and service ontology". This would indicate a degree of formal semantic and semantic richness that is neither required nor found in any of the current approaches. Thus, it seems to be more suitable to use the term "descriptive languages for products and services" in the future.

Currently, the prevailing architecture for such descriptive languages is the combination of

- classes,
- an attribute library,
- and class-specific attribute lists which contain references to characteristic attributes for the respective class (cf. eClass e.V. 2001, p. 30). Class-specific attribute lists support more precise product descriptions without an inflationary amount of classes. In addition to that, they allow parametric search (cf. eClass e.V. 2001, p. 23).

There was an approach to add attributes to UNSPSC, called “Universal Content Extended Classification System” (UCEC) (cf. Fensel et al. 2001, p. 57), which later merged into the “ECCMA Global Attribute Schema” (EGAS) (cf. Electronic Commerce Code Management Association 2001). As of today, though, there are no attribute lists available for the current UNSPSC version.

A suitable language architecture, however, is just a precondition for a good descriptive language. Equally important is its content, i.e. how well the available vocabulary and the expressive power of the language satisfy the linguistic requirements of business transactions. Two major indicators for this are the coverage of concepts and the semantic precision. Recent research has shown that one major problem of the development and maintenance of descriptive languages for products and services is the high linguistic dynamic in markets (cf. Hepp 2003, pp. 104-135): New products or business concepts evolve quickly and require new product classes. Existing products change and demand for new attributes.

Besides the percentage of real-world business concepts (e.g. common products, materials, or services) contained in the vocabulary of a descriptive language, the quality of attribute assignment influences its overall quality. This is a major challenge, as the following example demonstrates: The current version of eCl@ss contains more than 24,000 classes. If it takes a skilled standards engineer two hours to compose and verify a specific list of attributes for each class, and if he can work on this task eight hours per day, he will have to work on that for 6,000 days, or more than 16 years, seven days a week without a single day off. This indicates that the real challenge is in the attribute assignment.

Except for the raw number of elements (see Table 1), there are no general indicators available which help assess the content quality of descriptive languages for products and services. The mere number of elements, however, is a poor indicator of the actual quality of a descriptive language with attribute lists.

	eCl@ss 4.1	eCl@ss 5.0	eOTD 04-19-2003	UNSPSC V6 03-15-2003
Product and service classes	15315 (all levels)	24814 (all levels)	60474 (all levels)	19778
Total number of attributes in the attribute library	5504	3667	33160 (only 20829 used in at least one attribute list)	n/a

Table 1: Total Number of Elements in Product Classification Systems

This paper describes new quantitative measurements, which can be employed to assess the quality of existing systems and point to specific shortcomings. Besides the usage shown in this paper, standards bodies could use those new metrics to

- monitor the development of content quality,
- assess the amount of resources necessary to eliminate the shortcomings,
- rank content maintenance alternatives,
- and possibly motivate industry groups to help improve currently weak segments of the standard.

As the availability of a comprehensive descriptive language for products and services is crucial for an efficient market, the findings might also serve as arguments in favor of public funding.

The structure of this paper is as following. Section 2 shows how the distribution of product and service classes along the top-level hierarchy reveals the actual size of the vocabulary for specific segments. Section 3 describes a new approach for the analysis of attribute list quality based on their degree of specificity. This research yields two indicators that help monitor the progress in the development of descriptive languages for products and services. Section 4 summarizes the findings and implications.

2 Distribution of Classes along the Hierarchy

When buying an ordinary English dictionary, the mere number of words contained might be a suitable criterion for quality. A dictionary with 3,000 words will, in most cases, contain only a very basic vocabulary and thus support only a less specific mode of expression in comparison to an alternative dictionary with 60,000 words. Unfortunately, this rationale cannot be applied to descriptive languages for products and services. The reason is that the producers of dictionaries devote remarkable resources in order to keep the vocabulary balanced. This is not necessarily true for the bodies that develop descriptive languages. In fact, one can observe that a few categories (more precisely: clusters of meaning) contain far more entries than the average, and they grow faster, too. There are at least two mechanisms that fuel such uneven content development. First of all, it happens that organizations (e.g. industry associations, governmental bodies, or major corporations) join an existing approach. Consequently, they either contribute their proprietary classifications or help adding entries according to *their* urgent needs. Second, detecting the need for additions and updates of the language requires feedback from current or potential *users*. If, for example, an available descriptive language is attractive to the automotive industry because of its strong content in the respective cluster of meaning, it is far more probable that the body in charge will receive feedback about missing entries or proposals for improvement regarding automotive components. In contrast to this, the descriptive language will not be used in industries whose terminology needs are insufficiently covered. The consequence is that there will be less feedback on how to improve the weak categories. We will see below that uneven coverage and growth is a serious issue in current descriptive languages for products and services.

As said, most generic product classification schemes are built around a hierarchy. This hierarchy can be used for the analysis of the distribution of entries among clusters of meaning (cf. Hepp 2003, p. 141). The methodology employed is as following:

1. Create a list of top-level sections of the hierarchy.
2. Count all classes (entries) belonging to each top-level section.
3. Sort them by descending number of entries.

This approach was employed for the analysis of

- eCl@ss 4.1,
- eCl@ss 5.0,
- the ECCMA Open Technical Dictionary (eOTD, release retrieved April 4, 2003),
- and the UNSPSC Unified Version 6.0315 (retrieved March 15, 2003).

The distribution for eCl@ss 5.0 is shown in Figure 1. Respective distribution diagrams for the other systems can be found in (Hepp 2003, pp. 153, 166, and 177-178). It should be mentioned that the numbers for each top-level section include the top-level node itself and nodes on all subordinate levels.

This analysis confirms the assumption stated in the introduction. All tested classification systems have a few very huge categories. More than 20 % of the entries in eCl@ss and eOTD belong to the biggest single top-level section. Whilst the absolute percentage is already substantial, the lack of balance becomes even more evident when compared to the mean percentage (which is equal to the reciprocal value of the number of top-level categories). For example, eOTD is subdivided into 79 top-level categories, which would, given a completely balanced distribution, hold 1/79 or 1.3 % of the product classes. In reality, the biggest section contains 24.48 % of the total vocabulary. The three biggest categories altogether make up almost 40 % of the entries. UNSPSC proves to be more balanced with regard to its branches, with only 29.96 % in the three biggest categories. The findings are summarized in Table 2.

Even a perfect descriptive language for products and services will not have top-level sections that are completely equal in size, because different clusters of meaning can require a different semantic precision and thus a varying number of classes. This, however, neither explains nor justifies the observed magnitude of imbalance.

	eCl@ss 4.1	**eCl@ss 5.0**	**eOTD 04-19-2003**	**UNSPSC V6 03-15-2003**
Product classes	15315 (all levels)	24814 (all levels)	60474 (all levels)	19778 (all levels)
Number of top-level categories	22	25	79	55
Mean (percentage)	696 (4.5 %)	992 (4 %)	765 (1.3 %)	360 (1.8 %)
Amount of nodes in the biggest top-level category	3594	5303	14802	2447
Percentage	23.47 %	21.37 %	24.48 %	12.37 %
Amount of nodes in the three biggest top-level categories	6788	9818	24082	5860
Percentage	44.32 %	39.57 %	39.82 %	29.96 %

Table 2: Distribution of Classes along the Hierarchy

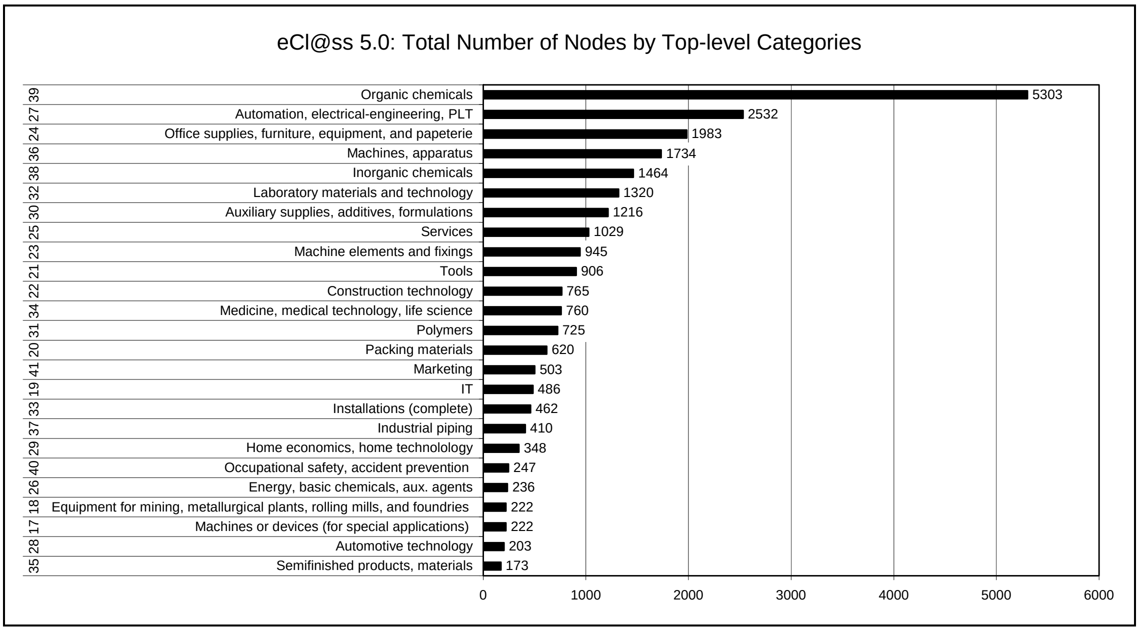

Figure 1: Distribution of Classes along the Hierarchy in eCl@ss 5.0

Besides the differences in content distribution among eCl@ss, UN/SPSC, and eOTD it might be relevant to determine the percentage of entries contained in all of the three. This could be used to assess the potential benefit of merging multiple approaches. However, the intersection of the three systems cannot be determined without comparing entry by entry manually. As an alternative, strengths and weak-nesses in content can be analyzed using sample baskets of goods and services.

Respective experiments with regard to commodities and office supplies are described in (Hepp 2003, pp. 142-142, 155, 164, and 175) and support the hypothesis that current classification systems lack frequently required products and services.

3 Quality of Attribute Assignment

A first approach to measure the quality of and progress in class-attribute assignment is counting the number of *node specific* attribute lists (cf. Hepp 2003, pp. 141 and 149-150). The need for this differentiation results out of the situation that usually a small set of generic attributes exists that is assigned to all (or a huge percentage) of the classes. In the context of eCl@ss, such unspecific attribute lists containing only a set of five generic attributes are named “base attribute lists”, as opposed to “standard attribute lists” (cf. eClass e.V. 2001, p. 30). OTTO and BECKMANN pointed out that “specific attribute list” would be a better term for the class-specific lists (see Otto / Beckmann 2001, p. 353). In eOTD, such base attributes and base attribute lists also exist, but are not explicitly labeled as such (see Hepp 2003, p. 171).

Thus, when measuring the progress with regard to attribute lists, the first step is to ignore such attribute lists that contain only generic attributes. In the context of this paper, an attribute is considered a base attribute when it is contained in more than 75 % of the attribute lists. The eCl@ss data model includes a field (“mkbsa”) that indicates a class-specific attribute set (“mkbsa = 2”). However, instead of just counting respectively marked classes, the real attribute list data was used. This revealed that there are four classes that indicate a specific attribute list in the aforementioned flag, but do not have specific attributes in their attribute list. This explains the difference between 7913 and 7917 specific attribute lists in Table 3.

The described procedure yields the percentage of classes with specific attribute lists. Table 3 shows the respective findings. The analysis yields that both eCl@ss 5.0 and eOTD provide specific attribute lists for only about 30 % of the nodes (cf. Hepp 2003, pp. 171-176). It is important to note that this percentage is based on the *total* number of classes on all levels, not just end nodes (leafs). This approach was chosen notwithstanding the fact that the current eCl@ss architecture defines attributes only on the leaf level (eClass e.V. 2001, p. 23), because there exist e-business scenarios that would benefit from fully functional superordinate concepts, i.e., generic terms with attributes. Relative to the number of leafs (20379) eCl@ss 5.0 has specific attribute lists for 38.83 % of the entries. The attribute assignment in the eOTD suffers from a lack of structure and consistency (see Hepp 2003, pp. 171-172): For example, the attribute “Brand Name” (EGAI 019543) is, without any observable rationale, assigned to only 17 of the 60474 classes.

	eCl@ss 4.1	eCl@ss 5.0	eOTD 04-19-2003
Number of node-specific attribute lists	6507	7913 (7917, see text)	19927 (quality issues, see text)
Percentage (based on the *total* number of product classes)	42.5 %	31.9 %	33 %

Table 3: Number and Percentage of Class-specific Attribute Lists

This basic indicator can be improved by analyzing the degree of specificity of the attribute lists. The fundamental idea is that an attribute that is used very frequently is generally less specific than an attribute used for only a few product categories. In the original form, an attribute list is considered specific as soon as it contains a single attribute that is used in less than 75 % of the attribute lists. The approach described in here consists of two steps: First, the *Semantic Weight* for each attribute in the attribute library is determined. In a second step, the *Semantic Value* for each single attribute list is computed by adding the semantic weights of all attributes contained. The semantic value for classes without an attribute list is by definition equal to zero.

Semantic Weight of Attributes: For each attribute

A_i with i = 1, … , *Number of Attributes*

in the attribute library, count the number of entries in the class-attribute relation. This yields the number of occurrences of attribute A_i. Thus, each attribute A_i in the attribute library receives a semantic weight SW_i that is equal to the reciprocal value of its usage frequency in a given release of the classification system (this idea resembles basic concepts in information and communication theory).

$$SW_i = \frac{1}{Number\,of\,Attribute\,Lists\,Containing\;A_i}$$

It is important to note that this is not a property of the respective attribute alone, but reflects its usage in a given classification system. The uneven distribution of classes as described in section 2 and the fact that node specific attribute lists do not yet exist for almost 70 % of the classes affect the absolute semantic weights.

A base attribute will have a semantic weight of

$$\frac{1}{\alpha * Number\,Of\,Attribute\,Lists} \quad \text{with } 1 \geq \alpha \geq 0.75$$

The value α reflects the percentage of attribute lists that contain this base attribute. It results from the definition of a base attribute as outlined above.

A very specific attribute used only in one single attribute list has a semantic weight of 1. Attributes in the attribute library that are not used in any attribute list should be simply ignored, because no meaningful value can be determined.

Table 4 summarizes the distribution of attribute usage frequency in eCl@ss 5.0. The analysis reveals four clusters, which are remarkably sharply separated. Note that the following percentages are based on the number of product classes (24814), not on the number of attribute lists (20379):

- Five attributes are assigned to more than 75 % of the classes. These attributes are identical to the base attributes
- 44 attributes are assigned to 25 – 28 % of the classes.
- 1942 attributes are assigned to less than 2 % of the classes, but used more than once.
- 1676 attributes are contained in exactly one attribute list.

Those attributes contained in only one attribute list deserve special attention, as their rare usage might point to duplicates in the attribute library. In some cases, the uneven distribution of classes along the hierarchy can explain such rarely used attributes.

Percentage of Classes with this Attribute	Number of Attributes	Share of the Attribute Library
> 75 %	5	0.14 %
29...74 %	none	
25...28 %	44	1.20 %
3...24 %	none	
< 2 % (but contained in more than one attribute list)	1942	52.96 %
contained in exactly one attribute list	1676	45.70 %

Table 4: eCl@ss 5.0: Clusters of Attribute Usage Frequency in Attribute Lists

Semantic Value of Attribute Lists: Now, for each product or service class C_j in the standard with an attribute set S_j, we sum up the semantic weights of all contained attributes.

This yields the semantic value SV_j for the Class C_j with j=1, … , *Number of Classes*

$$SV_j = \sum SWA_i \mid A_i \in S_j$$

The fundamental rationale is that more attributes mean a higher semantic specificity of the class, but very frequently used attributes add less semantic than specific attributes.

SV_j is an indicator for the semantic specificity of the class C_j. The higher SV_j, the more distinct is the respective attribute list from that of any other class.

It is important to note that the semantic value is not an absolute measurement, because it is influenced by the size and structure of the attribute library. For example, a badly structured attribute library with duplicate entries for identical properties will increase the semantic values. The major gain is not the value itself, but its *distribution properties* with regard to the descriptive language as a whole. Table 5 shows the distribution properties of the semantic values in eCl@ss 5.0. For this table, only the semantic values of the 20379 attribute lists were used. The 4435 (24814 classes – 20379 attribute lists) semantic values of zero for classes without an attribute lists were ignored.

Mean	0.17994013
Min	0.00019964
Max	87.88928283
Median	0.00025091
Mode	0.00025091

Table 5: eCl@ss 5.0: Distribution Properties of the Semantic Values (only classes with attribute lists)

A cluster analysis of the raw data reveals seven levels of semantic values (see Table 6):

- 4435 classes with no attribute lists and thus a semantic value of zero,
- 12466 classes with only base attributes in their attribute lists,
- 6344 classes with a limited number of branch specific, but very frequently used attributes. This group contains mostly chemicals.
- 1028 classes with a semantic value above 0.008, but less than 1. These can be considered as specific attribute lists with standardized and frequently used attributes.
- 437 classes with a semantic value greater or equal 1, but less than 10.
- 87 classes with a semantic value between 10 and 24.
- 17 classes with a semantic value greater than 24.

The two sections with the highest semantic values contain either classes with many singular attributes, or with a vast number of standardized attributes. One might suspect that the maximal value of 87.9 is mainly due to singular attributes. In fact, however, the single class with this high semantic value ("Bottom globevalve", primary key AAD661001, eCl@ss code 37-01-02-60) has 266 (!) attributes.

Semantic Value	Number of Classes	Remarks
24 – 88	17	
10 – 24	87	
1.0 – 9.99999999	437	
0.008 – 0.99999999	1028	
0.00026 – 0.00713688	6344	mostly chemicals
0.0002 – 0.00025095	12466	classes with base attribute lists and three defect base attribute lists
0	4435	classes without attribute lists

Table 6: eCl@ss 5.0: Semantic Value Clusters

4 Conclusion

The research described in this paper reveals that the impressive number of elements in both eCl@ss and eOTD obscure the very broad range of content quality in the diverse categories. Both standards have almost 40 % of their elements in the three biggest branches. Two potential mechanisms that have fostered such uneven development were identified.

Furthermore, it could be demonstrated that of the enormous task of assigning attributes to classes, only a minimal part has already been achieved. Of all current approaches, eCl@ss seems to be the leading descriptive language, but is far from being a comprehensive descriptive language for products and services.

The new metrics developed in this paper could be employed by standards bodies to monitor the development of content quality. The resulting values can help detect duplicate entries in the attribute library or inconsistent attribute assignment.

The findings could also be taken as an indicator in favor of attribute inheritance, which has the potential to ease attribute list maintenance and improve the consistency of attribute assignment.

References

Agrawal, R. / R. Srikant (2001): On integrating catalogs, in: Proceedings of the tenth international World Wide Web conference on World Wide Web (WWW10), Hong Kong, May 1-5, 2001, pp. 603-612.

Baron, J. P. / Shaw M. J. / Bailey, A. D. (2000): Web-based e-catalog systems in B2B procurement, in: Communications of the ACM, Vol. 43, No. 5, 2000, pp. 93-100.

eClass e.V. (2001): eClass - Leitfaden für Anwender, Alpha-Informationsgesellschaft mbH, Lampertheim 2001.

Electronic Commerce Code Management Association (2001): Content Europe Becomes ECCMA's Newest Charter Sponsor. URL: http://www.eccma.org/download/press_release, 2003. (Retrieved 04-21-2003)

Fensel, D. (2001): Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, Springer, Berlin etc. 2001.

Fensel, D. / Ding, Y. / Omelayenko, B. et al. (2001): Product Data Integration in B2B E-commerce, in: IEEE Intelligent Systems (Special Issue on Intelligent E-Business), Vol. 16, No. 4, 2001, pp. 54-59.

Guarino, N. (1998): Formal Ontology and Information Systems, in: Proceedings of the 1st International Conference on Formal Ontologies in Information Systems (FOIS'98), Trento, Italy, June 6-8, 1998, pp. 3-15.

Hepp, M. (2003): Güterklassifikation als semantisches Standardisierungsproblem, Deutscher Universitäts-Verlag, Wiesbaden 2003. [Zugl.: Diss., Würzburg, 2003]

Leukel, J. / Schmitz, V. / Dorloff, F.-D. (2002): Exchange of Catalog Data in B2B Relationships - Analysis and Improvement, in: Proceedings of IADIS International Conference WWW/Internet 2002 (ICWI 2002), Lisbon, Portugal, November 13-15, 2002, pp. 403-410.

Otto, B. / Beckmann, H. (2001): Klassifizierung und Austausch von Produktdaten auf elektronischen Marktplätzen, in: Wirtschaftsinformatik, Vol. 43, No. 4, 2001, pp. 351-361.

Stonebraker, M. / Hellerstein, J. M. (2001): Content Integration for E-Business, in: ACM SIGMOD 2001, Vol. 30, No. 2, ACM Press, New York, NY, USA 2001, pp. 552-560.