

# Content Metrics for Products and Services Categorization Standards

Martin Hepp<sup>1,2</sup>, Joerg Leukel<sup>3</sup>, Volker Schmitz<sup>3</sup>

<sup>1</sup>Florida Gulf Coast University, Fort Myers, FL, USA

<sup>2</sup>Digital Enterprise Research Institute (DERI), Innsbruck, Austria

<sup>3</sup>University of Duisburg-Essen, Essen, Germany

mhepp@computer.org, {joerg.leukel, volker.schmitz}@uni-essen.de

## Abstract

*Products and services categorization standards, such as eCI@ss, eOTD, RosettaNet Technical Dictionary (RNTD), or UNSPSC, play a major role for the automation of content integration tasks, because they provide a consensual vocabulary that can be used for the tagging of product-related data along the various stages of the product life cycle. Eventually, the quality and usefulness of a given categorization standard is determined by its content, especially the coverage of concepts in the respective application domain, its structure and semantic consistency, and the level of detail provided.*

*This paper proposes metrics for the content quality of products and services categorization standards, and applies those metrics to four prominent examples. It can be shown that there are significant differences between eCI@ss, eOTD, RNTD, and UNSPSC, which should both influence the choice of a standard for a specific business purpose and the maintenance strategies for the standards themselves.*

## 1. Introduction

Product-related data is exchanged frequently between multiple information systems, both inside the same organization and across company borders. The most obvious example is the exchange of product catalogs between vendors and electronic marketplaces, but equally important is the flow of itemized invoicing data for corporate spend analysis in global organizations. In most areas of application, the data from multiple sources must be aligned with a given target categorization, e.g. a taxonomy of cost categories or a catalog structure. This problem is referred to as content integration, and the content to be integrated is to a huge extent operational data and highly volatile [1]. The resulting time constraints require automated processing, which in turn requires machine-accessible semantics in the source data. The

common approach to provide such machine-accessible semantics is tagging individual data sets with references to entries in a standardized vocabulary of products and services terminology. Such collections of consensual concepts for the communication about products and services have been subject to much research in diverse research communities, e.g. under the label “ontologies” in the knowledge representation and data management field [2], with specific focus on catalog-data integration [3-6], and as “product classification standards” (PCS) [7, 8] or “product schema” [9] in the e-commerce arena. Also, “descriptive languages for products and services” has been proposed [10]. Within this paper, we refer to such standardized vocabularies for products and services terminology as *Products and Services Categorization Standards* (PSCS).

Many researchers have worked on the task of integrating two standards by finding similar concepts and establishing mappings between them, e.g. [8] or [11]. In our opinion, however, it is much more important that there exists at least one standardized vocabulary of vendor-independent domain concepts, and that this offers sufficient coverage and level of detail, than is a mapping between alternative standards. Very surprising is that the vast majority of previous work takes the existence of such categorization standards for granted and treats the most prominent approaches eCI@ss, ECCMA Open Technical Dictionary (eOTD), UNSPSC, or the RosettaNet Technical Dictionary (RNTD) as an externally given solution to the non-trivial requirement of sufficient coverage and detail.

Additionally, both the problem of selecting the right PSCS for a given business scenario and its representational needs, and the task of creating and maintaining a complex PSCS have been neglected so far. Those challenges require quantitative metrics for the content quality of a given PSCS. Only this allows for the comparison of multiple standards with regard to their quality.

In this paper we propose a comprehensive set of metrics for the content quality of products and services categorization standards, and show by means of comprehensive data analysis that those metrics

- (1) properly reflect the characteristics of a given standard and
- (2) yield findings that are useful for both the selection and assessment of appropriateness of a given PSCS by potential users, and for the maintenance and quality control of this standard by the responsible standards body.

Except for [10, 12], we do not know of any in-depth analysis of the content quality of PSCS. The empirical study by Fairchild and Vuyst analyzes the concepts of standardized PSCS [13], but describes only some characteristics of UNSPSC on a very high level of abstraction. Similar work to ours can be found in the ontology community in [14]; they propose metrics for the structural properties of RDF-S schemas for the Semantic Web, but include only one product-related schema in their analysis of 28 schemas.

The structure of this paper is as follows. In section 2 we propose metrics for the evaluation of products and services categorization standards. Section 3 briefly summarizes our findings gained by the application of those metrics to multiple releases of eCI@ss, the eOTD, RNTD, and the UNSPSC.

## 2. Content metrics

In the following, we present four main groups of metrics that address

- (1) the size, growth, and maintenance performance,
- (2) the degree of balance, hierarchical order, and breadth of coverage,
- (3) the size and expressiveness of the property library, and
- (4) the specificity of property assignment in class-wise property lists.

### 2.1. Size and growth

The metrics in this section reflect the size and pace of growth of a given PSCS by comparing *multiple releases of the same PSCS* with regard to the number of products and services classes, and relating the amount of new or modified elements to the amount of time passed between two release dates.

Those metrics show the amount of common concepts in the standard, i.e. those that reflect some degree of domain consensus. The metrics do not take into account the coverage of concepts of a specific

application domain. Measuring the growth and the maintenance work for a given PSCS per period of time indicates the amount of feedback received from the application domain and the “bandwidth” and delay of the standardization process, whichever is the limiting factor.

#### 2.1.1. Number of products and services classes

**Definition of the metric:** For each release of a specific PSCS we count the overall number of products and services classes. For hierarchically organized standards, we include intermediate nodes on all levels of the hierarchy. Then, we determine (1) the number of new and (2) the number of modified elements, i.e. such concepts that existed in the previous release but have now a new version number due to changes in the definition of the concept.

**Rationale:** This metric reflects the vocabulary size, i.e. the number of generic products and services concepts in the respective PSCS, and how this changes over time. It also shows the degree of change dynamics between any two subsequent releases, which is important for standards users, as it helps determine a suitable strategy to cope with release changes. Modified elements often require manual checking whether the existing class assignments are still valid.

**2.1.2. Speed of growth.** Due to the ongoing innovation in the products and services domain, standards bodies have to create new categories for new types of goods. The actual amount of new categories is constrained by at least two limitations: The amount of input received from the market side and the speed of processing such input.

**Definition of the metric:** For each release change of a given PSCS, we determine the amount of (1) new and (2) modified classes (if there is a hierarchical order: on any intermediate level) and divide it by the number of months passed since the two release dates.

**Rationale:** For a good coverage of concepts needed in the domain, any PSCS requires timely and complete feedback about missing entries from the user community, and a streamlined standardization process that makes respective new elements available in a timely manner.

### 2.2. Metrics for hierarchical order and balanced content

Most PSCS include a hierarchy of all products and services classes. This can be used to partition the total number of classes into the respective top-level sections and draw conclusions about the distribution along the

hierarchy. We can also use this approach for the analysis of how the distribution of classes develops over time, in order to see whether a given PSCS is getting more balanced or whether the degree of imbalance increases, and in which areas the content is actually being improved.

The resulting data is interesting, because it

1. reveals the degree of balance among the different categories and
2. shows the most populated categories and thus the domain focus of a PSCS.

Obviously, those metrics cannot be applied to standards that do not contain at least some form of hierarchical order.

### 2.2.1. Number of classes per top-level section

**Definition of the metric:** For each release of a given PSCS, we determine the total number of classes per each top-level, i.e. all descendents plus the top-level category itself. For the most recent version of the respective PSCS, the results can be summarized in a bar chart listing all top-level categories ordered by descending number of classes in this category.

**Rationale:** Many PSCS were created by merging existing standards from specific domains (eCI@ss: sourcing needs of the chemical industry; eOTD: NATO procurement). The mere numbers of categories often used for standards marketing obscure the true coverage in the various sections, because a few highly populated sections, resulting from the bulk import of sometimes very specific concepts, often contribute to a large amount of the total number of concepts.

**2.2.2. Services vs. products.** Categorization standards for goods can contain concepts for products, for services, or both. The mere existence of services categories, however, does not reveal the actual amount of services categories as compared to products.

**Definition of the metric:** We count the total number of services concepts (on all levels) based on the description of the first level of the hierarchy and relate them to the total number of concepts (on all levels). This approach does not take into account services that are hidden in a deeper level of the hierarchy, but the later can only be found by manually counting each single entry, which is unfeasible.

**Rationale:** The services domain differs from the representation of tangible products, e.g. because the fulfillment is bound to properties of the service customer, especially with regard to location and time. Also, there might be industries where, due to their high volume, services are of special interest for spend analysis. It thus makes sense to analyze the percentage of services categories in the total amount of categories.

### 2.2.3. Distribution properties of the number of classes per top-level section

**Definition of the metric:** We determine the distribution parameters for the data gained in section 2.2.1, i.e. the minimal value, maximal value, mean, median, first quartile, third quartile, interquartile range, standard deviation, and the coefficient of variation.

**Rationale:** These metrics show how the distribution of classes along the categories developed over time, in order to see whether a given PSCS is getting more balanced or whether the degree of imbalance increases. Also, since the coefficient of variation can be used to compare distributions that have a different mean, it is a good indicator for the comparison of multiple PSCS.

**2.2.4. Percentage of content in the most populated top-level categories.** In many PSCS, we find a few huge top-level categories, often reflecting bulk contributions from previous industry-specific efforts. The following metric sheds light on this issue.

**Definition of the metric:** For the current release of a given PSCS and based on the data gained in section 2.2.1, we determine the percentage of concepts contained in (1) the most populated and (2) in the three most populated top-level categories.

**Rationale:** For horizontal products and services standards, this reveals whether the standard is a true horizontal approach or horizontal just with regard to the existence of top-level categories, but focused quite vertically at the more detailed level. A true horizontal standard requires not only the existence of top-level categories for a broad range of concepts but also actual content in the deeper branches of all categories.

**2.2.5. Size of the most populated category vs. median of all top-level categories.** It is not only worthwhile to know the percentage of concepts in the biggest category, but also the order of magnitude of the imbalance between the biggest category and the median.

**Definition of the metric:** For the most recent version of a PSCS, we divide the number of elements in the most populated top-level category by the median of all categories.

**Rationale:** This metric reveals the order of magnitude of the number of concepts in the most populated top-level category as compared to the median. The bigger this ratio, the more is the content of the standard dominated by one single category.

**2.2.6. Number of descendents per superordinate category.** Besides the distribution of categories among the top-level sections, it is useful to see how the degree

of detail varies among the various levels of the hierarchy.

**Definition of the metric:** For each level of the hierarchy individually, we count the number of *direct* descendents per superordinate node, and determine the minimal value, maximal value, mean, median, standard deviation, and the coefficient of variation for the resulting data.

**Rationale:** This metric reveals how the degree of detail (i.e., the number sub-concepts) varies among the levels of the hierarchy.

## 2.3. Property library

Many PSCS include a library of standardized product properties. Those can be used to describe product instances with a far greater level of detail and allow parametric search.

### 2.3.1. Number of properties

**Definition of the metric:** For each release of a given PSCS, we count the total number of properties in the property library.

**Rationale:** The size of the property library reflects the amount of concepts for properties in the given standard. However, it can be suspected that redundancy is a big problem with regard to properties, because the often distributed development of PSCS makes it very likely that redundant properties are created when the existence of an equivalent property is not realized due to different terminological conventions. In its current stage, this is a rather raw metric, as it does not indicate the amount of consolidation work (e.g. the deletion of redundant properties).

**2.3.2. Enumerative data types.** Product properties (e.g. “disk diameter”) can either refer to a standard data type (e.g. integer, float, ...), often in combination with a unit of measurement (e.g. “inches”), or to a set of symbols reflecting valid concepts. The second form of data typing is usually referred to as enumerative data types, because the lexical space is an explicit set or list of usually a quite limited amount of items.

**Definition of the metric:** We count all properties in the property library that are assigned at least one enumerative data value and relate the number of those properties to the total amount of properties.

**Rationale:** It is highly desirable to have properly defined lexical spaces for all properties and thus enumerative data types for such properties that cannot be unambiguously represented using standard data types. However, we can often observe that such

property definitions are incomplete (e.g. defined as any alphanumeric sequence of less than 30 characters). This impedes automatic interpretation of property values.

## 2.4. Quality of class-specific property sets

Many PSCS contain a property-class relation that assigns necessary or recommended properties from the property library to individual products and services classes. This tells a standards user the suitable properties for the description of an item of the respective products or services category.

Unfortunately, the quality and specificity of those property-class assignments varies significantly. On one hand, there is usually a small set of very generic properties assigned to any (or almost any) class. Property lists containing just such generic properties add little to the description of a category. On the other hand, it happens that property lists hold a huge number of arbitrarily selected and often redundant properties.

A first approach to measure the quality of and progress in class-property assignment is to count the number of class-specific property lists. In the context of this paper, a property is considered a generic property when it is contained in more than 75 % of the property lists, and a property list is considered specific as soon as it holds one single specific (i.e. not generic) property.

### 2.4.1. Percentage of classes with specific property lists

**Definition of the metric:** We count all products and services categories that contain at least one specific property in their property list. Even if a given PSCS assigns properties only at the leaf level (i.e. no properties are assigned to intermediate nodes in the hierarchy), we compute, for reasons of comparability, the percentage based on the total number of concepts.

**Rationale:** Only the amount of specific property assignments indicates the amount of progress in the creation of fully-fledged products and services concepts. As an extension, this metric could be applied to each top-level category in order to identify those top-level categories that actually contain a high amount of specific property lists.

### 2.4.2. Property usage in property lists

**Definition of the metric:** For each concept that has a specific property list, we count the number of properties in this list and determine the minimal value, maximal value, mean, median, standard deviation, and coefficient of variation.

**Rationale:** Property lists should contain all necessary properties, but not a wild collection of any usable property, because this makes automated processing of product data difficult, as elements of the same type might be described using different properties. Additionally, a huge variation in the amount of properties per each good category indicates only partial progress in the creation of property assignments.

**2.4.3. Semantic weight and semantic value.** The basic metrics listed above can be improved by taking into account the degree of specificity of the property lists. The fundamental idea is that a property being used very frequently is generally less specific than a property assigned to only a few categories. In the metrics above, a property list is considered specific as soon as it contains a single property that is used in less than 75 % of all property lists. The extended approach described in this section consists of two steps: First, the *Semantic Weight* for each property in the property library is determined. In a second step, the *Semantic Value* for each single property list is computed by adding the semantic weights of all properties contained. The semantic value for classes without a property list is by definition equal to zero.

**Semantic weight of properties:** For each property  $P_i$  with  $i = 1, \dots$ , *Number of Properties*

in the property library, we count the number of entries in the class-property relation. This yields the number of occurrences of property  $P_i$ . Then, each property  $P_i$  in the property library receives a semantic weight  $SW_i$  that is equal to the reciprocal value of its usage frequency in a given release of the PSCS (this idea resembles basic concepts in information and communication theory).

$$SW_i = \frac{1}{\text{Number Of Property Lists Containing } P_i}$$

It is important to note that this is not a characteristic of the respective property alone, but reflects its usage in a given PSCS. The uneven distribution of classes and the fact that node specific property lists do not yet exist for a huge portion of the classes influence the absolute semantic weights.

A base property will have a semantic weight of

$$\frac{1}{\alpha * \text{Number Of Property Lists}}$$

with  $1 \geq \alpha \geq 0.75$

The value  $\alpha$  reflects the percentage of property lists that actually contain this base property. Its range results from the definition of a base property as above.

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

**Semantic value of property lists:** Now, for each product or service class  $C_j$  in the PSCS having a property set  $S_j$ , we sum up the semantic weights of all contained properties. This yields the semantic value  $SV_j$  for each Class  $C_j$  with  $j=1, \dots$ , *Number of Classes*

$$SV_j = \sum_{P_i \in S_j} SW_i / P_i \in S_j$$

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

$SV_j$  is an indicator for the semantic specificity of the class  $C_j$ . The higher  $SV_j$ , the more distinct is the respective property 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 property library. For example, a badly structured property 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 PSCS as a whole.

As an attempt to take into account the size of the property library and penalize overly big property collections with lots of redundant entries, the raw value  $SV_j$  should be divided by the number of properties.

### 3. Results

The application of the proposed content metrics reveals significant limitations of current releases of all major PSCS. Due to limited space, we can only present a small selection of our findings in this paper.

1. Even though eCl@ss, eOTD, and UNSPSC are regarded as horizontal categorization standards, they are quite vertically focused and have the majority of their classes in a few top-level categories. All three have more than 30% of all entries in three large sections and thus only a small partition of their 25 (eCl@ss), 55 (UNSPSC), or 79 (eOTD) top-level categories.

2. Both eCl@ss and UNSPSC undergo continuous improvement with an average of more than 200 new classes per month. On the other hand, eOTD had less than one new class per month in 2004, and this despite its wide coverage. It is hardly possible that there is no need for new classes in 79 categories. RNTD has also received only minimal additions with a mean of 1.3

new classes per month for the last five releases. For us this points to either lack of user feedback, lack of users, insufficient maintenance procedures, or any combination of these.

3. UNSPSC has by far the largest percentage of services categories, which might be due to the wide usage for spend analysis purposes. In absolute numbers, however, eOTD and eCI@ss offer also a remarkable amount.

4. With regard to the structure of its taxonomy, UNSPSC is very well balanced from the top-level down to its leaves. At the leaf level, the coefficient of variation of the amount of classes is only 100%. eCI@ss has a 56% higher variation and is obviously somewhat incomplete. Half of the 3<sup>rd</sup> level categories contain only two or less leaf-classes (median of 2).

5. RNTD has specific property lists for all of its classes, as compared to only 43% (eCI@ss) and 35% (eOTD). In other words, more than 2/3 of all eOTD classes and more than half of all eCI@ss classes are currently without specific property lists.

6. Classes in the middle of the distribution have a surprisingly similar amount of properties (44...47) in their property lists. The coefficient of variation is very similar among eCI@ss, eOTD, and RNTD, and with around 40% quite low. This seems to be the number of properties that are both manageable and sufficient for the description of products or services.

7. All PSCS have many properties that are used with only one or two classes. This can point either to redundancy, to the "arbitrary" creation of property lists on demand, or a combination of both.

8. Both eCI@ss and eOTD show an enormous spread with regard to the semantic value of their property lists. The coefficient of variation is as high as 523% (eCI@ss) and 432% (eOTD). This very well reflects our observation that both have huge differences in the quality of the property assignment. RNTD has a mean about 33 (eCI@ss) to 100 times (eOTD) of the size. Those orders of magnitude are very much compatible with our manual observations.

As a summary, the proposed metrics are quite successful in revealing weaknesses with regard to the content quality of current PSCS. While we applied the metrics to the full population data, they can also be used to assess selected top-level sections that are of interest for a specific application. Corporations can for example compare the semantic values and the amount of maintenance work for their industry segments among multiple PSCS. This will prevent investment into such standards that neither cover existing representational needs nor show convincing efforts of improvement. Standards bodies should use those metrics as indicators for their own progress.

## References

- [1] M. Stonebraker and J. M. Hellerstein, "Content Integration for E-Business," Proc. of the ACM SIGMOD 2001, Santa Barbara (CA), USA, 2001.
- [2] L. Obrst, R. E. Wray, and H. Liu, "Ontological Engineering for B2B E-Commerce," Proc. of the International Conference on Formal Ontology in Information Systems (FOIS'01), Ogunquit, Maine, USA, 2001.
- [3] O. Corcho and A. Gómez-Pérez, "Solving Integration Problems of E-commerce Standards and Initiatives through Ontological Mappings," Proc. of the Workshop on E-Business and Intelligent Web at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001), Seattle, USA, 2001.
- [4] D. Fensel, Y. Ding, B. Omelayenko, E. Schulten, G. Botquin, M. Brown, and A. Flett, "Product Data Integration in B2B E-Commerce," *IEEE Intelligent Systems*, vol. 16, pp. 54-59, 2001.
- [5] D. Fensel, D. L. McGuinness, E. Schulten, W. K. Ng, E.-P. Lim, and G. Yan, "Ontologies and Electronic Commerce," *IEEE Intelligent Systems*, vol. 16, pp. 8-14, 2001.
- [6] B. Omelayenko, "Ontology Integration Tasks in Business-to-Business E-commerce," Proc. of the Fourteenth International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, June 4-7, 2001, Budapest, Hungary, 2001.
- [7] J. Leukel, V. Schmitz, and F.-D. Dorloff, "A Modeling Approach for Product Classification Systems," Proc. of the 13th International Workshop on Database and Expert Systems Applications (DEXA'02), Aix-en-Provence, France, 2002.
- [8] E. Schulten, H. Akkermans, G. Botquin, M. Dörr, N. Guarino, N. Lopes, and N. Sadeh, "The E-Commerce Product Classification Challenge," *IEEE Intelligent Systems*, vol. 16, pp. 86-89, 2001.
- [9] G. Yan, W. K. Ng, and E.-P. Lim, "Product Schema Integration for Electronic Commerce - A Synonym Comparison Approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, pp. 583-598, 2002.
- [10] M. Hepp, "Measuring the Quality of Descriptive Languages for Products and Services," Proc. of the Multikonferenz Wirtschaftsinformatik 2004, Essen, 2004.
- [11] D. Beneventano, F. Guerra, S. Magnani, and M. Vincini, "A Web Service based framework for the semantic mapping amongst product classification," *Journal of Electronic Commerce Research*, vol. 5, pp. 114-127, 2004.
- [12] M. Hepp, *Güterklassifikation als semantisches Standardisierungsproblem*, Deutscher Universitäts-Verlag, 2003.
- [13] A. M. Fairchild and B. de Vuyst, "Coding Standards Benefiting Product and Service Information in E-Commerce," Proc. of the 35th Annual Hawaii International Conference on System Sciences (HICSS-35), pp. 258b, 2002.
- [14] A. Magkanaraki, S. Alexaki, V. Christophides, and D. Plexousakis, "Benchmarking RDF Schemas for the Semantic Web," Proc. of the First International Semantic Web Conference (ISWC2002), pp. 132-146, 2002.