Alignment of Product Master Data

Thomas Kirsche, Gerhard Baumann, Anja Schanzenberger

GfK Marketing Services
Nordwestring 101, 90319 Nuremberg, Germany
{thomas.kirsche|gerhard.baumann|anja.schanzenberger}@gfk.de

Abstract: Market research draws a coherent picture of the market based on extensive observations of sales acts from numerous data sources. As the data sources refer to the products sold all over the world in different formats and with different keys, the data needs to be aligned to a common product master. The subtle strategies for a large-scale product data alignment, as well as the key structure of comprehensive product master database are explained.

1 Introduction

Big turnover and good profit is good news for industry and retail, but this information alone is nothing worth if it cannot be related to product segments of a market. For example, the trade partners want to learn about the latest brand shares of digital still cameras segmented by mega pixels. Product-related market segments can be built upon brands, product groups (e.g. digital still cameras), or properties of the products (e.g. mega pixels). Thus, top-notch market research is driven by comprehensive and detailed product information.

Ranked among world's top 10 market research companies, GfK Group delivers a wide range of business information on markets for corporate decision makers. The clients of GfK Marketing Services, a business unit within GfK Group, are based in the consumer goods manufacturing industry. GfK Marketing Services' product is information about the markets of consumer electronics, information technology/telecom, domestic appliances, photo and do-it-yourself, as well as related areas and services. The product is periodically created out of a wealth of sales activity, taken as sales records from tens of thousands retail outlets in 60+ countries. The underlying product master database has a value of its own and is even offered as a separate by-product (www.encodex.com).

Apart from facts like price and sales units, the sales information of a retailer includes product identification information. This is typically the brand and model text/model number of the product sold. Because retailers do not (yet) use a common standard for product descriptions the various different product master data must be aligned for the purposes of market report creation.

In this paper, we describe how we align the product masters of the numerous different sources with a continuous effort, as new products arrive every single day in the market

(see section 0). In section 0 the key elements of the product master database are outlined. Section 4 briefly touches upon our plans for the future.

2 Product Master Data

In this section, a brief overview is given of the product master data structure in StarTrack, GfK Marketing Services' data production system [RK05] [ALT97]. As it turns out, the product data world is not flat but a multi-level hierarchy to accommodate the needs of reporting and data alignment.

2.1 Product Data Elements

Due to space limitations, it is impossible to outline the complete product data model of StarTrack, so the main product data elements are named (Figure 1).

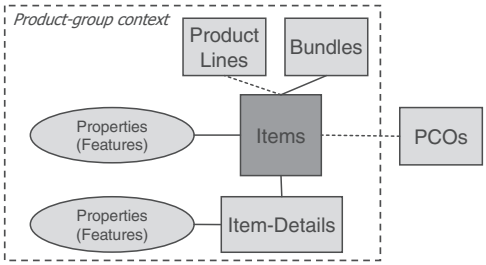

Figure 1: Data entities of the StarTrack product master

In the center of all product data is what is called an item. An item belongs to a product group which defines the realm of product characteristics. For example, the Canon EOS 350D, an 8 mega pixel SLR camera, belongs to the product group of digital still cameras. The items of this product group are said to be coded against mega pixels and lots of other properties, commonly referred to as features. Items may have country-specific properties like availability, market introduction dates and more, as an item is typically an abstraction of what a customer can actually buy. The items are linked to item-keys, synonymously called item-details, or details, for short. Typically, the details represent the EAN code of an item, for example one EAN code for one color. Again these properties might be country-specific.

Product lines act like a product family. The members of this family have a common set of characteristics which is why StarTrack enforces identical feature values within the defined product line features and family members, respectively. For example, the Canon EOS series might be such a product family within digital still cameras. Bundles are sets of items or details that are sold together. Typically, SLR camera bodies are sold with lenses in a kit.

As the product group forms the precise classification context of its items, it is a given that items where the product group is (yet) unknown cannot be coded. To overcome this limitation, so-called 'partly classified objects' (PCOs) provide a space to record the item characteristics although the features must be manually selected in advance, item by item.

All items, item-details, PCOs, product lines have a unique key, called ObjectId. The properties/features are addressed with ObjectId and country code. The majority of the data elements are kept as versions so that a previous state of the information can be recovered.

The StarTrack product master does not exactly follow a standard like UNSPSC or eCl@ss (for a description and methodological sound comparison see [He03]). GfK acts, however, as a voting member in bodies like EAN.UCC Global Product Classification Task Group (GPCTG) and makes sure that best ideas of standards fertilize market research and vice versa.

2.2 Key Systems

Once the product data of different sources has been aligned, the alignment decision is stored as key-pairs. In addition to the key system of the product master, retailer-specific key systems, so-called instore codes, and global key system like the EAN/JAN or UPC are used most often. This framework can easily be extended to accommodate key-systems specific to manufacturers or application domains, like the ISBN number for books. For data alignment purposes it matters whether a key system is exclusively used with one data source or shared. In the latter case, unknown items might be automatically aligned based on a previous alignment decision for a different data source. As Figure 1 suggests, keys are considered as a part of the product master data.

2.3 Replacement Operator

A common problem in data quality is items are recorded twice in the database but they actually refer to the same product. One reason for duplicates within a product group creation is the same product has been sold with different model texts, or varying prefixes or different suffixes. For example, Canon sells the same product in Europe as “EOS 350D” and in the US as “Digital Rebel XT”. Cross-product groups duplicates are a phenomena appearing along with service integration of the products: Smart phones could be seen as mobile phones or personal digital assistants (PDAs). To overcome the human error of duplicate creation, the ‘replacement’ operator allows merging two objects into one. Using various options, feature values can or cannot be merged, and the already aligned sales data will be aggregated. Replacements fully honor the complex object structure of items, details, and product-lines.

3 Alignment Strategies

The business process of data alignment at GfK Marketing Services is triggered by the new arrival of sales tracking data. A retailer’s sales data has embedded references to the products sold, but no link to the StarTrack product master. The goal of data alignment, synonymously called identification, is to link all relevant sales records with the corresponding product master data at GfK. Figure 2 depicts the overall process.

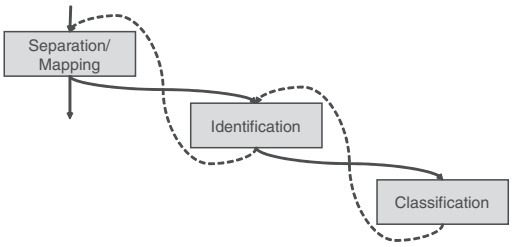

Figure 2: Data Alignment Process

As the majority of the product references in the tracking data usually have been previously processed, the known data is first mapped to the StarTrack product master by simple reference look-ups. Product references which are not known are separated, grouped into single instances and then pushed forward to identification. Again, the majority of retailer references can be aligned to existing product master records. The established links are equivalence or subclass-of relationships [CG01]. Then, the result of identification is sent back to separation/mapping. Retailer references that cannot be aligned due to missing product master entries trigger an extension of the product master data. This is the manual classification of new products. Subsequently, we highlight points of interest in the data alignment process.

3.1 Proposals

Unfortunately, the actual identification process is used to be a time-consuming, manual work. Thus, a core strategy is to support this task by a proposal system. Using smart search algorithms on the retailer data and GfK's master data, there has been considerable progress in creating proposals for alignments. The proposals are created by a step-wise extension of the search domain in the product master, while keeping it small enough to create meaningful results. That is, an initial search strategy is selected on the basis of the product group and the retailer. Some particular attributes, like brands or kernel number, are treated special. If a kernel number of the model text exists, it is the largest consecutive sequence of digits. For example, 350 is the kernel number of "EOS 350 D". Additionally, the price of a product is taken into account, so that the possible matches could only include products in a valid price range.

The proposal system yields the best match with a high confidence, but there is no way to directly use the proposals to replace manual identification, even if the proposal system returns exactly one result. The most important obstacle is the incompleteness of the product master itself. The best match could also be a wrong match, because the right match would be the link to a missing product master record. This record needs then to be created manually in classification.

3.2 Cross Alignment

A new product enters usually the market at different retailers but to the same time. As a consequence, the data alignment process encounters the equal unknown sales record at different retailers, i.e. in different formats, spellings, and with different keys. In fact it is hard to see that the different sales records relate to the same product.

GfK Marketing Services uses intelligent sorting and filtering mechanisms to show the semantic proximity of unrelated items. In addition to text patterns filters, a valuable set of filters relies on knowledge of previous alignments. This knowledge is used to determine a likely product group for unknown records, based on the fact that similar records have been aligned with this product group. For example, to a good extent, the unknown records can be filtered that will be most probably aligned with digital cameras.

If records actually have a common reference like the EAN code, the alignment of one record is pushed to all the other EAN codes, thereby identifying hundreds of sales records within seconds.

4 Conclusion

To ensure data quality, data alignment of product master data is of paramount importance in market research, if data from external sources is processed directly. GfK Marketing Services manages successfully to perform large-scale data alignments as part of their daily business. Developments underway will yield additional quality for the cross alignment process with a scoring model. Only if an alignment decision has been confirmed, thereby exceeding a minimum score, the alignment originally valid for a subset of data sources will become generally valid. Other research directions go into the supervision of the whole production process [SL04].

References

[ALT97] Albrecht, J.; Lehner, W.; Teschke, M.; Kirsche, T.: Building a Real Data Warehouse for Market Research. In (Hameurlain, A.; Tjoa, A. M. Eds.): Proc. 8th Int. Conf. on Database and Expert Systems Applications (DEXA '97), Toulouse, 1997. LNCS 1308: Springer, Berlin, 1997; pp. 651-656.

[CG01] Corcho, O.; Gómez-Pérez, A.: Solving integration problems of e-commerce standards and initiatives through ontological mappings. In: Proc. IJCAI'01 Workshop on Ontologies and Information Sharing, Seattle, 2001; pp. 131-140.

[He03] Hepp, M.: Güterklassifikation als semantisches Standardisierungsproblem. DUV, Wiesbaden 2003.

[RK05] Ruf, T.; Kirsche T.: Data Refinement in a Market Research Applications' Data Production Process. In (Härder, T.; Lehner, W. Eds.): Data Management in a Connected World. LNCS 3551: Springer, Berlin, 2005; pp. 293-314.

[SL04] Schanzenberger, A.; Lawrence, D.R.: Automated Supervision of Data Production – Managing the Creation of Statistical Reports on Periodic Data. In: (Meersman, R.; Tari, Z. Eds.): On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE; Proc. OTM Confederated Int. Conf. CoopIS, DOA, and ODBASE (CoopIS 2004), Larnaca, 2004. LNCS 3290: Springer, Berlin, 2004; pp. 194-208.