Using BMEcat Catalogs as a Lever for Product Master Data on the Semantic Web

Alex Stolz, Bene Rodriguez-Castro, and Martin Hepp

E-Business and Web Science Research Group, Universität der Bundeswehr München
Werner-Heisenberg-Weg 39, D-85577 Neubiberg, Germany
{alex.stolz,bene.rodriguez}@ebusiness-unibw.org, mhepp@computer.org

Abstract. To date, the automatic exchange of product information between business partners in a value chain is typically done using Business-to-Business (B2B) catalog standards such as EDIFACT, cXML, or BMEcat. At the same time, the Web of Data, in particular the GoodRelations vocabulary, offers the necessary means to publish highly-structured product data in a machine-readable format. The advantage of the publication of rich product descriptions can be manifold, including better integration and exchange of information between Web applications, high-quality data along the various stages of the value chain, or the opportunity to support more precise and more effective searches. In this paper, we (1) stress the importance of rich product master data for e-commerce on the Semantic Web, and (2) present a tool to convert BMEcat XML data sources into an RDF-based data model anchored in the GoodRelations vocabulary. The benefits of our proposal are tested using product data collected from a set of 2500+ online retailers of varying sizes and domains.

1 Introduction

Online shopping has experienced significant growth during the last decade. Preliminary estimates of retail e-commerce sales in the US show an increase of 17.3% from the third quarter of 2011 to the third quarter of 2012, while they grew to almost five times 2003 levels, totaling 5.2 percent (57 billion dollars) of the entire retail sales market [15]. These recent statistics indicate a large body of different-sized online stores ranging from major retailers like Amazon, BestBuy or Sears to small web shops offering only tens or hundreds of products. Hence it comes as no surprise that instances of popular commodities are offered by a fairly large number of shopping sites. Many of those online shops maintain databases where they can store information and data to describe their goods. Nonetheless, for site-owners it proves difficult to get hold of rich and high-quality product data for all of their items over time, especially if their specifications originate from product catalogs by different manufacturers. Large size retailers might obtain this information in a semi-automated fashion via some form of catalog exchange format. However, small shop owners might have to enter products and feature data manually. This scenario produces repeated definitions of the same product features, but mainly with incomplete, inconsistent and outdated information

Table 1. Comparison of product features between manufacturers and retailers

Manufacturer Product Features		**Retailer Product Features**		**Coverage**[1]
Samsung LED TV ES6300	89	15	amazon.de	28.09%
		39	notebooksbilliger.de	
		22	conrad-electronics.de	
		24	voelkner.de	
Siemens Kettle TW86103	25	10	amazon.de	23.64%
		22	redcoon.de	
		4	quickshopping.de	
		13	elektro-artikel-shop.de	
Suunto M5 Running Pack	33	12	amazon.de	49%
		3	sportscheck.com	
		1	otto.de	
		15	klepsoo.com	
		8	tictactime.de	

across various online retailers. Little and inaccurate information about products ultimately hampers the effective matchmaking of products.

Another source of product data for commodities are their manufacturers. These compile and maintain specifications of all of their products. Typically, their product catalogs are managed in Product Information Management (PIM) systems that can export content to different types of media, e.g. via electronic product catalogs as seen on many manufacturer sites or printed catalogs. PIM systems host essential and core product data also known as *product master data*.

Table 1 presents a simple illustration of the situation using the example of three random products. The table compares the number of features provided by the goods' manufacturers with the features found at a large leading online retailer and other online merchants of various sizes selected arbitrarily via the "Shopping" service of Google Germany[2]. Unless otherwise specified, by "features" we mean structured product specifications (i.e datasheets in tabular form published on the shop pages) without taking into account product pictures, product name and product description. It can be seen that the product data provided across the different sources vary significantly.

To date, product master data is typically passed along the value chain using Business to Business (B2B) channels based on Electronic Data Interchange (EDI) standards such as BMEcat (catalog from the German Federal Association for Materials Management, Purchasing and Logistics[3]) [12]. Such standards can significantly help to improve the automatic exchange of data. However, trading partners still have to negotiate and set up information channels bilaterally, which prevents them from establishing ad-hoc business relationships and raises the barriers for potential business partners that either do not have the means or the money to connect via imposed B2B standards. Similarly, end users, who

[1] "Coverage" = Ratio of average number of retailer features and manufacturer features
[2] `http://www.google.de/shopping/`
[3] English for "Bundesverband Materialwirtschaft, Einkauf und Logistik e.V. (BME)"

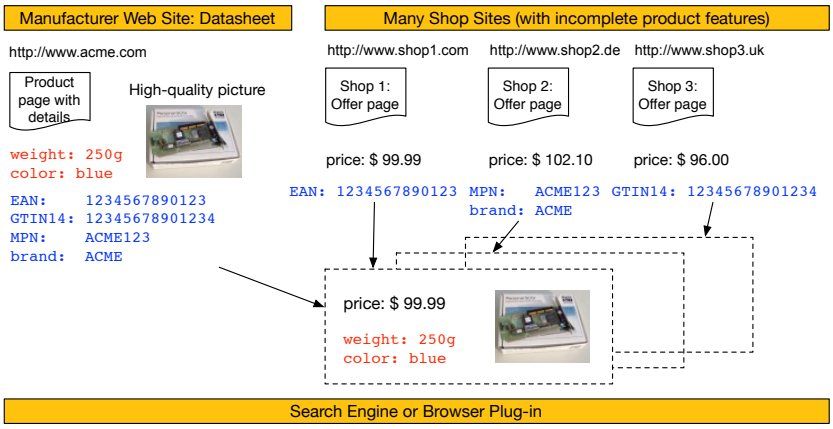

Fig. 1. Lever of manufacturer product master data using *strong identifiers*

could benefit from enterprise data liberalization by facing better search and matchmaking services for products, are neglected [4].

An approach to tackle this issue is to publish rich product master data straight from the Product Information Management (PIM) systems of manufacturers on the Web of Data, so that it can be electronically consumed by other merchants intending to trade these goods. Under this premise, retailers and web shop owners could then rely on widely used product *strong identifiers* such as European/International Article Number (EAN), Global Trade Item Number (GTIN), or Manufacturer Part Number (MPN), to leverage this rich data straight from manufacturers. Fig. 1 illustrates an example of this approach, where three different online merchants benefit from product descriptions and features as published by the manufacturer relying on the corresponding product strong identifier. Each online merchant can then use this rich manufacturer information to augment and personalize their own offering of the product in question.

In this paper, we propose to use the BMEcat XML standard as the starting point to make highly structured product feature data available on the Web of Data. We describe a conceptual mapping and the implementation of a respective software tool for automatically converting BMEcat documents into RDF data based on the GoodRelations vocabulary [9]. This is attractive, because most PIM software applications can export content to BMEcat. With our approach, a single tool can nicely bring the wealth of data from established B2B environments to the Web of Data. Our proposal can manifest at Web scale and is suitable for every PIM system or catalog management software that can create BMEcat XML product data, which holds for about 82% of all of such software systems that we are aware of, as surveyed in [17]. Furthermore, it can minimize the proliferation of repeated, incomplete, or outdated definitions of the same product master data across various online retailers; by means of simplifying the consumption of authoritative product master data from manufacturers by any size of online retailer. It is also expected as a result that the use of structured data in terms of

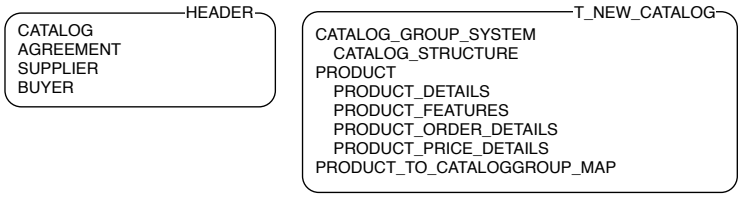

Fig. 2. BMEcat 2005 skeleton

the GoodRelations vocabulary by manufacturers and online retailers will bring additional benefits derived from being part of the Web of Data, such as Search Engine Optimization (SEO) in the form of rich snippets[4], or the possibility of better articulating the value proposition of products on the Web.

To test our proposal, we converted a representative real-world BMEcat catalog of two well-known manufacturers and analyzed whether the results validate as correct RDF/XML datasets grounded in the GoodRelations ontology. Additionally, we identified examples that illustrate the problem scenario described relying on structured data collected from 2500+ online shops together with their product offerings. Our tests allowed us to confirm the immediate benefits and impact that adopting our approach can bring to both manufacturers and retailers.

2 Conversion from BMEcat to GoodRelations

In this section, we first introduce background information on the BMEcat standard and the GoodRelations vocabulary. Then we present key alignments and challenges underlying the conversion from BMEcat to GoodRelations.

2.1 Background

Both BMEcat and GoodRelations share the goal to facilitate e-commerce transactions and product data exchange between business parties.

BMEcat. BMEcat is a powerful XML standard for the exchange of electronic product catalogs between suppliers and purchasing companies in B2B settings. The current release is BMEcat 2005 [12], a largely downwards-compatible update of BMEcat 1.2. The most notable improvements over previous versions are the support of external catalogs and multiple languages, and the consistent renaming of the ambiguous term *ARTICLE* to *PRODUCT*. Fig. 2 presents a high-level view of the document structure for the transmission of a catalog using BMEcat 2005. A valid BMEcat document comprises a header and a payload section:

– The header defines global settings such as defaults for currency, eligible regions or catalog language, and specifies seller and buyer parties involved in the transaction. It further may state the agreement or contract that the document is based on. The default values specified in the document header can be overwritten by values defined at product instance level in the document.

[4] `http://support.google.com/webmasters/bin/answer.py?hl=en&answer=99170`

– The payload section consists of a product data section and data related to classification standards (e.g. eCl@ss, UNSPSC)[5] or vendor-specific catalog group systems. Product data sections consist of product-related information, feature data, price details, and order details. The element name of the payload part determines the transaction type and can be one of *T_NEW_CATALOG* (new catalog), *T_UPDATE_PRODUCTS* (update of product data), and *T_UPDATE_PRICES* (update of price data).

GoodRelations. GoodRelations [9] is a light-weight vocabulary (ontology, schema, data dictionary) for e-commerce on the Semantic Web. Its expressivity is targeted at the description of an offer and its related entities, i.e. the description of relationships between business entity, offer, and product or service. The ontology provides basic support for the most frequently used properties and individuals in offering descriptions, such as product details, prices, and terms and conditions. The GoodRelations ontology allows to extend products (*gr:SomeItems* or *gr:Individual*) with product models (*gr:ProductOrServiceModel*), or datasheets, that can contribute detailed product information like product features. For that purpose, it provides a fully-fledged meta-model for expressing quantitative and qualitative product properties in OWL. In addition, to further categorize products and to describe them more precisely, GoodRelations allows to extend products and product models with classes and features of comprehensive product classification standards (e.g. eClassOWL [6] or the Product Types Ontology[6]).

To refer to GoodRelations elements in the remainder of this paper, we will use the commonly accepted namespace prefix *gr:*, which can be employed to shorten the full URI of the ontology, i.e. `http://purl.org/goodrelations/v1#name` becomes *gr:name*. Accordingly, we will omit any namespace declarations in text and tabular descriptions.

2.2 Alignments

In the following, we outline correspondences between elements of BMEcat and GoodRelations and propose a mapping between the BMEcat XML format and the GoodRelations vocabulary. Given their inherent overlap, a mapping between the models is reasonable with some exceptions that require special attention. We will highlight these cases, nonetheless we can not cover the full alignment here.

For the mapping between the two schemas the following aspects were considered: Company details (address, contact details, etc.), product offer details, catalog group structures, product features including links to media objects, and references to external product classification standards. Furthermore, multi-language descriptions in BMEcat are handled properly, namely by assigning corresponding language tags to RDF literals. An illustrative example of a catalog and its respective conversion is available online[7]. However, in the context of this

[5] `http://www.eclass.de/`, `http://www.unspsc.org/`
[6] `http://www.productontology.org/`
[7] `http://www.ebusiness-unibw.org/projects/bmecat2goodrelations/example/`

Table 2. Mapping of product details from BMEcat to GoodRelations

BMEcat	**GoodRelations**
PRODUCT	gr:Offering, gr:Individual/gr:SomeItems, gr:ProductOrServiceModel
SUPPLIER_PID type={*ean*, *gtin*}	gr:hasEAN_UCC-13, gr:hasGTIN-14
PRODUCT_DETAILS	
DESCRIPTION_SHORT lang={*en*, *de*, ...}	gr:name with language *en*, *de*, ...
DESCRIPTION_LONG lang={*en*, *de*, ...}	gr:description with language *en*, *de*, ...
INTERNATIONAL_PID type={*ean*, *gtin*}	gr:hasEAN_UCC-13, gr:hasGTIN-14
MANUFACTURER_PID	gr:hasMPN
MANUFACTURER_NAME	gr:hasManufacturer → gr:BusinessEntity → gr:name
PRODUCT_STATUS type={*new*, *used*, ...}	gr:condition

paper we focus solely on product model data. Also, we do not provide alignments for full classification standards that can be exchanged since BMEcat 2005, primarily because of the complexity and for legal reasons especially gaining in importance when converting licensed classification standards. Moreover, there already exist proposals that focus on the conversion and publication of product classification standards (e.g. eClassOWL [6]).

Product Details. At the center of the proposed alignments are product details and product-related business details. Table 2 shows the BMEcat-2005-compliant mapping for product-specific details. Table 2 adds an additional level of detail to the *PRODUCT* → *PRODUCT_DETAILS* structure introduced in Fig. 2. The element name highlighted in bold font face determines a new nesting level, e.g. *PRODUCT* consists of an attribute for the product identifier of the supplier and a sub-element *PRODUCT_DETAILS*. The elements discussed in the present context are all mapped to properties of product instances, product models and offers in GoodRelations. However, our main interest lies in the alignment to *gr:ProductOrServiceModel*. The product identifier can be mapped in two different ways, at product level or at product details level, whereby the second takes precedence over the other. Whether the European Article Number (EAN) or the Global Trade Item Number (GTIN) is mapped depends on the *type*-attribute supplied with the BMEcat element. Furthermore, the mapping at product level allows to specify the manufacturer part number, product name and description, and condition of the product. Depending on the language attribute supplied along with the *DESCRIPTION_SHORT* and *DESCRIPTION_LONG* elements in BMEcat 2005, multiple translations of product name and description can be obtained. Lastly, the manufacturer name is mapped to a little more complex pattern in GoodRelations, i.e. the value of *MANUFACTURER_NAME* maps to the name of the legal entity attached to the product model via *gr:hasManufacturer*.

Product Features. BMEcat allows to specify products using vendor-specific catalog groups and features, or to refer to classification systems with externally defined categories and features. The mapping of product classes and fea-

Table 3. Mapping of product features from BMEcat to GoodRelations

BMEcat	**GoodRelations**
PRODUCT_FEATURES	
REFERENCE_FEATURE_SYSTEM_NAME	referenced classification system identifier
REFERENCE_FEATURE_GROUP_ID	rdf:type (class id of classification system)
REFERENCE_FEATURE_GROUP_NAME	gr:category
FEATURE	
FNAME	rdfs:label and property name in GR
FDESCR	rdfs:comment
FVALUE	gr:hasValueFloat
FUNIT	gr:hasUnitOfMeasurement
FREF	feature id of referenced classification system, property name in GR context

Table 4. Mapping of a catalog group system in BMEcat to a *rdfs:subClassOf*-hierarchy

BMEcat	**GoodRelations**
CATALOG_GROUP_SYSTEM	
CATALOG_STRUCTURE	owl:Class
GROUP_ID	class name of owl:Class
GROUP_NAME lang={*en*, *de*, ...}	rdfs:label with language *en*, *de*, ...
GROUP_DESCRIPTION lang={*en*, *de*, ...}	rdfs:comment with language *en*, *de*, ...
PARENT_ID	rdfs:subClassOf (class id of superclass)

tures is shown in Table 3. The target GoodRelations property of the *REFERENCE_FEATURE_GROUP_NAME* element is *gr:category*. *REFERENCE_FEATURE_SYSTEM_NAME* (e.g. *ECLASS-5.1*) and *REFERENCE_FEATURE_GROUP_ID* have no direct mapping, rather a combination of them unambiguously determines the class identifier of a reference classification system (e.g. eClassOWL [6]). Likewise, the *FREF* element can be used together with *FVALUE* and an optional *FUNIT* element to specify a feature whose property is referenced externally. Otherwise, if no *FREF* is available for a feature, then the feature is defined locally. The *FUNIT* element can be used to discern property types in GoodRelations, i.e. to assign a quantitative object property to the product model in RDF if a value for *FUNIT* is given, otherwise a datatype property. The distinction will be addressed in more detail in Section 2.3.

Catalog Group Systems. Catalog groups are a means to further refine product descriptions. A catalog group system is mapped building up an *rdfs:subClassOf*-hierarchy based on the GenTax algorithm [10], which permits to create meaningful ontology classes for a specific context while at the same time preserving the original hierarchy, i.e. the catalog group taxonomy. Table 4 outlines the mapping of catalog groups in BMEcat to RDF. The hierarchy is determined by the group identifier of the catalog structure that refers to the identifier of its parent group.

Product and Catalog Group Map. In order to link catalog groups and products, BMEcat maps group identifiers with product identifiers using *PROD-*

UCT_TO_CATALOGGROUP_MAP. Accordingly, products in GoodRelations are assigned corresponding classes from the catalog group system, i.e. they are defined as instances (*rdf:type*) of classes derived from the catalog group hierarchy.

2.3 Design Decisions

In the following, we cover aspects of the conversion where the alignment of the two schemas turned out to be challenging.

Datatype versus Object Properties. OWL distinguishes between object properties and datatype properties [1]. The former category describes properties that link between individuals, whereas the latter links individuals with data values (literals), e.g. an entity with a numeric value or a textual description. The GoodRelations vocabulary further refines the categorization made by OWL by discerning qualitative and quantitative object properties. On the other side, BMEcat does not explicitly discriminate types of features, so features (*FEATURE*) typically consist of *FNAME*, *FVALUE* and, optionally, an *FUNIT* element. The presence of the *FUNIT* element helps to distinguish quantitative properties from datatype and qualitative properties, because quantitative values are determined by numeric values and units of measurements, e.g. *150 millimeters* or *1 bar*. Thus, any other feature is either a qualitative or a datatype property.

It is impossible to reliably discern qualitative properties and datatype properties in an automated way during conversion (e.g. are S, M, and L qualitative values describing garment sizes or rather simple literal values?), so we reserve this task for solving in the RDF world (potentially bringing in additional knowledge) and declare all such properties as datatype properties with a range of type string.

For those features whose values likely qualify as boolean values we provide a simple heuristic, i.e. if the feature value is one of "y", "n", "yes", "no", "true", or "false", then the property is assumed to be a boolean datatype property. Similarly, all rules that apply to properties also apply to their respective values, i.e. a quantitative property implies quantitative values, and so forth.

Float Value Ranges in Datatype Properties. Unlike GoodRelations, BMEcat does not allow to model range values by definition. There are two possibilities to model them in BMEcat, though. Either the BMEcat supplier defines two separate features, or the range values are encoded in the *FVALUE* element of the feature. The first option defines a feature for the lower range value and a feature for the upper range value, respectively. The downside of this approach is that two unrelated GoodRelations properties arise. The second alternative, i.e. range values encoded as single feature values, leads to invalid values (e.g. *gr:hasValueFloat "10-20"^^xsd:float*) when mapped to GoodRelations. For that reason, typical value patterns describing upper and lower ranges (like operating temperature of "5-40" degrees Celsius) are mapped to *gr:hasMinValueFloat* and *gr:hasMaxValueFloat* of quantitative values in GoodRelations. This approach, however, works only for common encoding patterns for range values in text.

Units of Measurement. BMEcat and GoodRelations recommend to use UN/CEFACT [14] common codes to describe units of measurement. In reality, though, it is common that suppliers of BMEcat catalogs export the unit of measurement codes as they are found in their PIM systems. Instead of adhering to the standard 3-letter code, they often provide different representations of unit symbols, e.g. *cm*, *centimeters*, etc. in place of CMT, which would be the correct UN/CEFACT code. This is inconvenient with regard to potential applications that should consume the data and compare products upon feature descriptions.

2.4 Implementation

The implementation of the logic behind the alignments to be presented herein resulted into the BMEcat2GoodRelations tool. BMEcat2GoodRelations is a portable command line Python application to facilitate the conversion of BMEcat XML files into their corresponding RDF representation anchored in the GoodRelations ontology for e-commerce. Due to the limited length of this paper, we refer readers to the project landing page hosting the open source code repository[8], where they can find a detailed overview of all the features of the converter, including a comprehensive user's guide.

3 Evaluation

To evaluate our proposal, we implemented two use cases that allowed us to produce a large quantity of product model data from BMEcat catalogs. We tested the two BMEcat conversions using standard validators for the Semantic Web, presented in Section 3.1. Then we compare the product models obtained from one of the BMEcat catalogs with products collected from Web shops through a focused Web crawl. Finally, we show the potential leverage of product master data from manufacturers with regard to products offered on the Web.

3.1 Validation of Use Cases

We tested our conversion using BMEcat files from two manufacturers, one in the domain of high-tech electronic components (Weidmüller Interface GmbH und Co. KG[9]), the other one a supplier of white goods (BSH Bosch und Siemens Hausgeräte GmbH[10]). In the case of Weidmüller, the conversion result is available online[11]. The products in the BSH catalog were classified according to eCl@ss 6.1, whereas Weidmüller provide their own proprietary catalog group system. This allowed us to validate the BMEcat converter comprehensively. Although the conversions completed without errors, still a few issues could be detected in each dataset that we will cover subsequently.

[8] `http://code.google.com/p/bmecat2goodrelations/`
[9] `http://www.weidmueller.com/`
[10] `http://www.bsh-group.com/`
[11] `http://catalog.weidmueller.com/semantic/sitemap.xml`

Table 5. Validation of BMEcat conversions

Validation	**BSH**	**Weidmüller**
BMEcat2GoodRelations converter	warnings: (a) wrong values where numeric values were expected; (b) non-standard unit codes detected	warnings: (a) non-standard unit codes detected
RDF Validator	valid. warning: invalid lexical value for literal	valid
W3C RDF Validation	valid	valid
Pellet	valid. warning: malformed xsd:float detected	valid
GoodRelations Validator	step 32 failed: non-compliance of float literal with xsd:float	valid

To validate the output of our conversion, we used publicly available online and offline validators. In addition to that, our converter prints helpful warning messages to the standard output. In summary, the converter was tested using the following validation steps: (1) BMEcat2GoodRelations converter output (including error and warning messages, if any), (2) RDF/XML syntax validity[12], (3) Pellet validation[13] for spotting semantic, logical inconsistencies, and (4) GoodRelations-specific compliance tests[14] to spot data model inconsistencies.

The converter has built-in check steps that detect common irregularities in the BMEcat data, such as wrong unit codes or invalid feature values. In Table 5, we list a number of warning messages that were output during the conversion of the BMEcat files, together with the validation results of the different validation tools. As shown in the table, the two conversions pass most validation checks, with a few data quality issues reported by some validators. In the BSH catalog for example, some fields that require floating point values contain non-numeric values like "/", "0.75/2.2", "3*16", or "34 x 28 x 33.5", which originates from improper values in the BMEcat. Another data quality problem reported is the usage of non-uniform codes for units of measurement, instead of adhering to the recommended 3-letter UN/CEFACT common codes (e.g. "MTR" for meters, "VLT" for Volt, etc.).

3.2 Missing Product Features on the E-Commerce Web of Data

Table 1 in the introduction showed how the number of features published by manufacturers does not always end up in the descriptions of the offerings published by online retailers. In this section, we elaborate on a complementary example that uses structured data on the Web of Data.

In addition to the manufacturer BMEcat files, we took a real dataset obtained from a focused crawl whereby we collected product data from 2629 shops. The dataset has a slight bias towards long-tail shops. Furthermore, the Web shops were not crawled entirely. Nonetheless, Fig. 3 illustrates the distribution of the product

[12] `http://www.rdfabout.com/demo/validator/`, `http://www.w3.org/RDF/Validator/`
[13] `http://clarkparsia.com/pellet/`
[14] `http://www.ebusiness-unibw.org/tools/goodrelations-validator/`

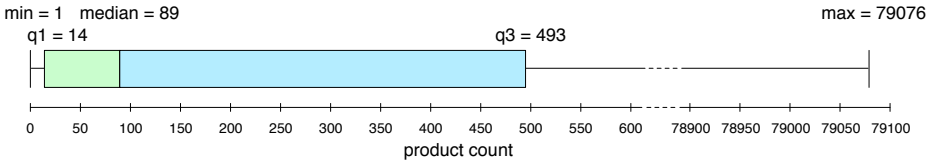

Fig. 3. Boxplot of distribution of product count across Web shops

count across shops for a snapshot of the crawl. To remove any potential bias caused by multiple definitions of the same product on different pages (because of non-canonical URIs containing query strings like `prod_id=1&sess_id=XYZ`), the boxplot was generated using the count of products with distinct EANs per shop. The upper quarter of shops offer more than 493 products according to Fig. 3. More interestingly, half of the shops offer less than 89 distinct products, whereas the majority of shops (three quarters) have less than 14 products. This could be explained either by the fact that several shops are rather small and provide only a limited set of offers, or by the non-comprehensive crawl of shop domains.

In Table 6, we complement the example given in the introduction with insights from our collected data. The products listed in the table represent product models from the BSH dataset and product instances from Web shops based on overlapping EANs. In the current dataset, there exist 95 of such matches based on EANs. The comparison of the amount of properties from the manufacturer with the number of properties from the retailers shows a significant gap. For instance, take the vacuum cleaner (German: *Bodenstaubsauger*) in row 2 of Table 6. It shows 30 product properties coming from the manufacturer and an average number of nine properties across the three shops that offer the product. Therefore, the properties in the shops only amount to a fraction (30%) of the properties available from the manufacturer. The relatively constant number of properties for product instances may be explained by the shop extensions that typically only express standard features like product name, GTIN, EAN, SKU, product weight and dimensions. Although this might to a certain extent explain the numbers, it does not change our premise that structured product master data is still lacking on the Web.

We collected all the data in an SPARQL-capable RDF store and extrapolated some statistics to substantiate the potential of our approach. The number of product models in the BSH was 1376 with an average count of 29 properties, while the Weidmüller BMEcat consisted of 32585 product models with 47 properties on average created by our converter. By contrast, the nearly 2.7 million product instances from the crawl only contain eleven properties on average.

3.3 Potential Leverage of Product Master Data on the Web

Table 6 from Section 3.2 confirmed the scenario presented in Table 1 in the introduction in the context of BSH product models and a sample of 2500+ online shops that provide structured data.

In this section, we present some specific examples of the number of online retailers that could readily benefit from leveraging our approach. To remain in

Table 6. Product features in BSH BMEcat and retailers publishing GoodRelations

BSH Product Features		Retailer Product Features		Coverage[15]
TW86103 Wasserkocher (EAN: 4242003535615)	25	10	marketplace.b2b-discount.de	40%
Bodenstaubsauger Beutel VS06G2410 2400 W (EAN: 4242003356364)	30	10	www.ay-versand.de	30%
		9	www.megashop-express.de	
		8	fairplaysport.tradoria-shop.at	
Mikrowelle HF25M5L2 Edelstahl (EAN: 4242003429303)	51	7	www.european-gate.com	13.73%

the scope of the use cases discussed, the examples are chosen from the BSH BMEcat products catalog, within the German e-commerce marketplace.

We chose to check for the number of shops offering products using a sample size of 90 random product EANs from BSH BMEcat. The sample size was selected based on a 95% confidence level and 10% confidence interval (margin of error), i.e. requiring a minimum of 90 samples given the population of 1376 products in the BMEcat. Using the sample of EANs, we then looked up the number of vendors that offer the products by entering the EAN in the search boxes on Amazon.de, Google Shopping Germany, and the German comparison shopping site preissuchmaschine.de[16]. This gave us a distribution of shops grouped by EAN as outlined in the boxplots in Fig. 4.

The numbers we got from this experiment were lower than expected. For example, the maximum number of sellers offering a specific product was 48. For half of the products that we tested at least 16 offers appeared in the price comparison search engine. In the Amazon.de and Google Shopping Germany marketplaces by comparison, the number of offers for a product among the sample of product EANs was even lower. We can think of various explanations for this, namely that the marketplace regulations try to limit competition among market participants and, more importantly, that adding products to the marketplace presents a barrier to smaller shop owners (in the case of Google Shopping, a shop is asked to upload product data using a populated product feed or an API). Furthermore, the small numbers may be due to (1) localized searches (*.de*-domain), (2) the fact that shops rarely populate their products with EAN identifiers, or (3) the type of products in our sample, in this case from the domain of white goods that are likely not that popular for being sold online. More precisely, unsupported small shop owners may not find it very attractive to sell dishwashers online given the effort involved for logistics.

To put Fig. 4 (boxplots) in perspective, we did a comparison with a more popular product, i.e. "Canon PowerShot A2300 schwarz" (with EAN "8714574578828"). We repeated the above searches with the same online services, but now using (a) the EAN of this digital camera and (b) the product name, suspecting that many retailers do not populate their products with EAN but use other strong identifiers instead. Amazon.de and preissuchmaschine.de constantly gave 45 and

[15] "Coverage" = Ratio of average number of retailer features and BSH features

[16] `http://www.amazon.de/`, `http://www.google.de/shopping/`, `http://www.preissuchmaschine.de/`

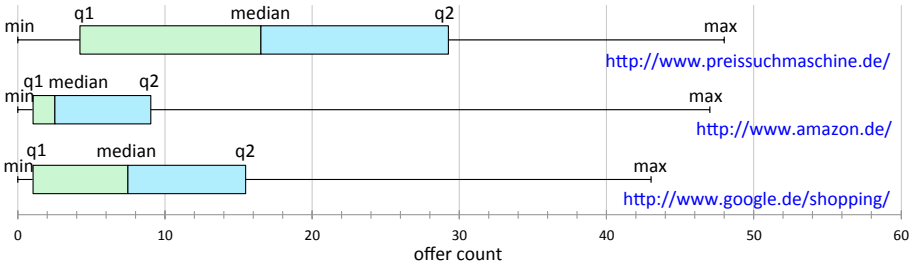

Fig. 4. Boxplots of distribution of shop offers per European Article Number (EAN)

233 results, respectively. Google Shopping Germany, however, returned only 4 results when searching by the EAN number, but 144 results for a search by product name. These results indicate that using a combination of different types of strong identifiers could leverage product master data on the Semantic Web.

4 Related Work

The rise of B2B e-commerce revealed a series of new information management challenges in the area of product data integration [5,13]. Separately, the gradual realization of the Semantic Web vision has motivated significant efforts aimed at representing existing e-commerce product related data and classification standards adopting open semantic technologies and data models [7,8,2].

Yet, in the context of managing *product master data* in particular, two previous solutions stand out [3,16] based on their similarities with respect to our problem scenario. The study in [3] presents a meta-model in OWL DLP (which expressivity profile lies between OWL 1 Lite and OWL 1 DL) as part of a semantic application framework that can provide semantic capabilities to a generic PIM system. On the other hand, [16] has developed an extension that allows lifting the data from existing relational databases of leading Master Data Management (MDM) systems into RDF format. This allows semantic interoperability across organizations' core data, applications and systems.

Both solutions share our reliance on Semantic Web technologies to facilitate product master data integration and consistency across separate data sources. However, there are several aspects where they deviate from our proposal as presented in the sections above, most notably: (a) Their scope focuses on closed corporate environments which may involve proprietary applications or standards rather than open technologies at the scale of an open Web of Data; and (b) being aimed at generic PIM and MDM systems, their level of abstraction is very broad, introducing additional degrees of separation with respect to the applicability to the problem scenario targeted by the BMEcat2GoodRelations converter tool.

In that sense, BMEcat2GoodRelations is to the best of our knowledge the only solution developed with open standards, readily available to both manufacturers and retailers to convert product master data from BMEcat into structured RDF data suitable for publication and consumption on the Web of Data.

5 Conclusions and Outlook

The proliferation of online retailers in recent years was accompanied by a growing number of products being offered on the Web. Such a substantial increase of online goods introduces new data management challenges. More specifically, it involves how information, in particular products, features or descriptions, can be processed by stakeholders along the product life cycle. Our experience after a survey of 2500+ different-sized online merchants indicates that in the current conditions product data suffers from incomplete, inconsistent or outdated information.

As a partial solution to mitigate the shortage of missing product master data in the context of e-commerce on the Web of Data, we propose the BMEcat2GoodRelations converter. This ready-to-use solution comes as a portable command line tool that converts product master data from BMEcat XML files into their corresponding OWL representation using GoodRelations. All interested merchants have then the possibility of electronically publishing and consuming this authoritative manufacturer data to enhance their product offerings relying on widely adopted product *strong identifiers* such as EAN, GTIN, or MPN.

We argue that the construction of a firm basis of product master data is the prerequisite for useful matchmaking scenarios. The data we collected and analyzed, provides enough evidence to motivate on the one hand a critical mass of manufacturers to release their product master data and on the other hand retailers to attach strong identifiers to their products. The immediate impact would be a huge lever for enriching online offers by product features and less effort to be put into data cleansing thanks to a gain in more high-quality data. Both factors would pave the way to more granular data analysis and search experience for organizations and individuals.

Acknowledgments. The authors would like to thank Mark Mattern, who provided a first mapping from BMEcat to GoodRelations as part of a master thesis supervised by Martin Hepp [11]. The work on this paper has been supported by the German Federal Ministry of Research (BMBF) by a grant under the KMU Innovativ program as part of the Intelligent Match project (FKZ 01IS10022B), and by the Eurostars program (within the EU 7th Framework Program) of the European Commission in the context of the Ontology-based Product Data Management (OPDM) project (FKZ 01QE1113D).

References

1. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A.: OWL Web Ontology Language Reference. Tech. rep., World Wide Web Consortium (2004), `http://www.w3.org/TR/owl-ref/`
2. Beneventano, D., Montanari, D.: Ontological Mappings of Product Catalogues. In: Shvaiko, P., Euzenat, J., Giunchiglia, F., Stuckenschmidt, H. (eds.) OM. CEUR Workshop Proceedings, vol. 431. CEUR-WS.org (2008)

3. Brunner, J.S., Ma, L., Wang, C., Zhang, L., Wolfson, D.C., Pan, Y., Srinivas, K.: Explorations in the Use of Semantic Web Technologies for Product Information Management. In: Proceedings of the 16th International Conference on World Wide Web. pp. 747–756. ACM, New York, NY, USA (2007)
4. Di Noia, T., Di Sciascio, E., Donini, F.M., Mongiello, M.: A System for Principled Matchmaking in an Electronic Marketplace. In: Proceedings of the 12th International Conference on World Wide Web. pp. 321–330. ACM, New York, NY, USA (2003)
5. Fensel, D., Ding, Y., Omelayenko, B., Schulten, E., Botquin, G., Brown, M., Flett, A.: Product Data Integration in B2B E-Commerce. IEEE Intelligent Systems 16(4), 54–59 (2001)
6. Hepp, M.: eClassOWL: A Fully-Fledged Products and Services Ontology in OWL. In: Poster Proceedings of the 4th International Semantic Web Conference. Galway, Ireland (2005)
7. Hepp, M.: Products and Services Ontologies: A Methodology for Deriving OWL Ontologies from Industrial Categorization Standards. International Journal on Semantic Web and Information Systems 2(1), 72–99 (2006)
8. Hepp, M.: ProdLight: A Lightweight Ontology for Product Description Based on Datatype Properties. In: Abramowicz, W. (ed.) Business Information Systems, Lecture Notes in Computer Science, vol. 4439, pp. 260–272. Springer, Heidelberg (2007)
9. Hepp, M.: GoodRelations: An Ontology for Describing Products and Services Offers on the Web. In: Gangemi, A., Euzenat, J. (eds.) Knowledge Engineering: Practice and Patterns, Lecture Notes in Computer Science, vol. 5268, pp. 329–346. Springer, Heidelberg (2008)
10. Hepp, M., De Bruijn, J.: GenTax: A Generic Methodology for Deriving OWL and RDF-S Ontologies from Hierarchical Classifications, Thesauri, and Inconsistent Taxonomies. In: Franconi, E., Kifer, M., May, W. (eds.) The Semantic Web: Research and Applications, Lecture Notes in Computer Science, vol. 4519, pp. 129–144. Springer, Heidelberg (2007)
11. Mattern, M.: Transforming BMEcat Catalogs into Semantic Web Annotation Data for Offerings. Master thesis, University of Innsbruck, Innsbruck, Austria (2009)
12. Schmitz, V., Leukel, J., Kelkar, O.: Specification BMEcat 2005. Bundesverband Materialwirtschaft, Einkauf und Logistik e.V., Frankfurt am Main, Germany (2005)
13. Stonebraker, M., Hellerstein, J.M.: Content Integration for E-Business. In: Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data. pp. 552–560. ACM, New York, NY, USA (2001)
14. United Nations Economic Commission for Europe (UNECE): Recommendation No. 20: Codes for Units of Measure Used in International Trade. UN/CEFACT Information Content Management Group (2006)
15. United States Census Bureau: Quarterly Retail E-Commerce Sales: 3rd Quarter 2012. U.S. Department of Commerce, Washington, DC, USA (2012)
16. Wang, X., Sun, X., Cao, F., Ma, L., Kanellos, N., Zhang, K., Pan, Y., Yu, Y.: SMDM: Enhancing Enterprise-Wide Master Data Management Using Semantic Web Technologies. Proc. VLDB Endow. 2(2), 1594–1597 (2009)
17. Weber, A.: Marktanalyse von Software für Produkt-Informations-Management (PIM). Bachelor thesis, Universität der Bundeswehr München, Neubiberg, Germany (2011)