Ontology-based Data Integration for Corporate Sustainability Information Systems

Lisa Madlberger
Vienna PhD School of Informatics
Vienna University of Technology
lisa.madlberger@tuwien.ac.at

Andreas Thöni
Information & Software Engineering Group
Vienna University of Technology
andreas.thoeni@tuwien.ac.at

Peter Wetz
Doctoral College Environmental Informatics
Vienna University of Technology
peter.wetz@tuwien.ac.at

Alexander Schatten
Sophisystems GmbH
alexander.schatten@sophisystems.com

A Min Tjoa
Information & Software Engineering Group
Vienna University of Technology
amin@ifs.tuwien.ac.at

ABSTRACT

Nowadays, the term sustainability gains more and more importance in corporate risk management and decision making. Corporate Sustainability Information Systems should support companies to analyse sustainability risks and provide relevant data in a practical manner. A major challenge in the domain of sustainability is to integrate environmental, social and economic data from different sources. This paper presents an approach for ontology-based data integration in the Corporate Sustainability domain preserving hierarchical relations to allow e.g. improved querying. Based on an existing sustainability framework defined in an XBRL-taxonomy, first a mechanism for the automatic generation of a domain-specific ontology is proposed. Finally, special data is queried and linked to the proposed ontology, serving as a working example for the suggested approach.

Categories and Subject Descriptors

H.4.2 [**Information Systems Applications**]: Types of Systems—*Decision support*; J.1 [**Administrative Data Processing**]: Business

General Terms

Design

Keywords

Information Integration, Corporate Sustainability, OWL, XBRL

iiWAS2013, 2-4 December, 2013, Vienna, Austria.

1. INTRODUCTION

In recent years, companies face increased pressure of consumers, non-governmental organizations (e.g. Greenpeace), governments and investors to take over responsibility for their impacts on the society and the environment. Compelling is the case of Nike, which faced an extensive consumer boycott after the New York Times and other newspapers published reports about abusive labour practices at some of its Indonesian suppliers in the early 1990s [21]. As a result, companies experience a business need to manage next to economic also environmental and social aspects, summarized under the caption "Corporate Sustainability". Corporate Sustainability Information Systems are a novice kind of information systems, which aim to support companies to analyse sustainability risks and to prevent negative consequences of incidents for the company and its reputation.

An information system monitoring sustainability risks has to include information from various heterogeneous sources to provide a comprehensive overview. Potential data to integrate includes energy or labour data from environmental or human resource systems [15], public events from news sources [11] or statistical information extracted from Linked Open Data (LOD) sources. Moreover, a large amount of companies already publish sustainability reports which could also serve as data input.

To establish a sound information base, these heterogeneous data sources have to be integrated. Within the Semantic Web, a machine readable version of the Web, ontologies are applied to overcome semantic ambiguity [18] caused by source heterogeneity. In the context of a Corporate Sustainability Information System, an ontology could enable semantic integration of input source data and a common understanding of concepts across different system components. However, up to now, no ontology has received broad recognition in the domain of Corporate Sustainability.

In this paper, we propose a top-down approach for the development of a Corporate Sustainability ontology and we show by an example how this ontology can be used to link sustainability-relevant data.

2. ONTOLOGY DEVELOPMENT

In this data integration scenario, the ontology to be developed mainly fulfills two tasks: (1) reduce semantic ambiguity of input sources by describing sources using a common vocabulary, and (2) structure input sources in a way, i.e. hierarchically, that reflects the business conception of Corporate Sustainability. Hierarchical structuring can be of significant importance since it enables flexible query formulation in order to aggregate data [23]. Additionally, hierarchies allow for data mappings on different categorization levels (see Section 3.1).

2.1 Requirements to the ontology

Therefore the ontology mainly has to fulfill the following requirements:

- Req. 1: The ontology should provide an intuitive representation of Corporate Sustainability concepts.
- Req. 2: The ontology should reflect the hierarchical structure of sustainability aspects (e.g. child labor as a subconcept of human rights).
- Req. 3: The ontology should allow a general mapping of sustainability sources to predefined concepts.

2.2 Top-down Ontology generation

Since on the one hand we face a broad range of potential very heterogeneous input sources (operative IT systems, news reports, sensor measurements, public statistics, etc.), but on the other hand a broad, relatively clearly defined business conceptualization of Corporate Sustainability, we propose a top-down approach generating an ontology from an existing business framework. By selecting an established standard the requirement of intuitive representation should be fulfilled as good as possible.

Several standards have been proposed to structure sustainability management (indicators) in corporations. Most popular is the standard GRI 3.1 defined by the Global Reporting Initiative (GRI), a global multi-stakeholder non-governmental organization [10]. GRI covers various criteria including economic, environmental and social ones such as corruption or child labour.

Recently, the GRI published an XBRL (eXtensive Business Reporting Language) taxonomy, with the goal to make sustainability data more standardized and exchangeable. XBRL, an XML-based data standard for publishing and exchanging business reports, was originally developed for reporting of financial information, nowadays XBRL-taxonomies are created also for other business areas, such as sustainability reporting.

As the GRI framework is established as the defacto standard in multinational companies with regard to Corporate Sustainability, it can be considered as an intuitive representation (*Requirement 1*). Therefore, we use the GRI-XBRL taxonomy as a basis to automatically generate a Corporate Sustainability ontology, which can be applied for data integration tasks in Corporate Sustainability Information Systems.

3. APPROACH

In order to define an ontology to be used for enterprise integration tasks in the Corporate Sustainability domain,

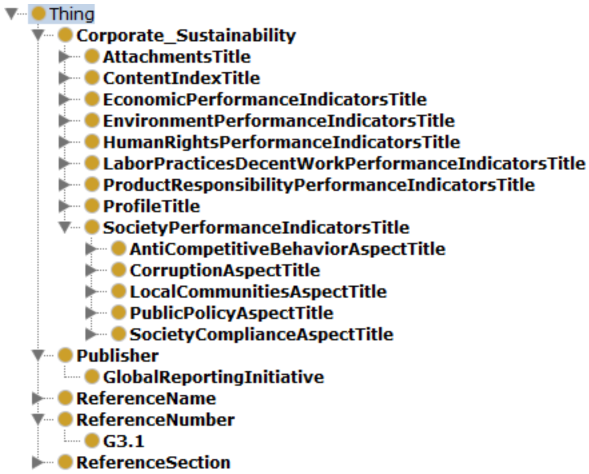

Figure 1: An extract of the generated Corporate Sustainability ontology based on the XBRL taxonomy.

we propose using the existing GRI-XBRL taxonomy as a starting point. The ontology-based sustainability data integration approach mainly requires three steps: (1) definition of a mapping between the XBRL-taxonomy and the target ontology, (2) definition of a transformation algorithm, (3) linking of input sources to the resulting ontology.

3.1 Mapping definition

Semantic integration requires a consistent understanding of the meaning of concepts. These can be modeled using the Web Ontology Language (OWL) that emerged as the norm for the description of ontologies in the Semantic Web [5].

Since the hierarchical structure of concepts is particularly important in our use case (*Requirement 2*), we define that XBRL-elements, the main concepts in an XBRL-taxonomy, correspond to OWL-classes in the resulting ontology. In an XBRL-taxonomy, relationships between elements are specified using XLinks. We use the links specifying "parent-child" relationships, to generate a class hierarchy using the "rdfs:subClassOf" property. We extract further information from the taxonomy, such as references, which define a particular section in a legal document, and indicator labels (see Table 1 for the complete mapping).

3.2 Transformation algorithm

Since both the source taxonomy as well as the target ontology, can be represented in XML-format, we use Extensible Stylesheet Language Transformations (XSLT), a language designed specifically for transformations of XML documents, for the transformation. The GRI-XBRL taxonomy consists of multiple files defining reporting elements, links between them, references to related documents and definitions. In essence, the algorithm first builds a class-structure from concepts defined in the core-taxonomy file and consequently integrates additional information from other files, which are connected to the core-elements via XLinks, e.g. references and labels. An exemplary class definition of the resulting ontology is presented in Listing 1.

Figure 1 shows an extract of the resulting ontology, that can be used for data integration tasks. Especially higher levels of this ontology can help to structure input data sources according to different sustainability subdomains. Based on

Table 1: Mapping of XBRL to OWL

XBRL-tag, xlink-connector	OWL
<element>	owl:Class
<presentationArc>, xlink:arcrole="http://www.xbrl.org/2003/arcrole/parent-child"	rdfs:subClassOf
<ref:Publisher>, xlink:role="http://www.xbrl.org/2003/role/reference"	csOnt:hasReferencePublisher
<ref:Name>, xlink:role="http://www.xbrl.org/2003/role/reference"	csOnt:hasReferenceName
<ref:Number>, xlink:role="http://www.xbrl.org/2003/role/reference"	csOnt:hasReferenceNumber
<ref: Section>, xlink:role="http://www.xbrl.org/2003/role/reference"	csOnt:hasReferenceSection
<label>, xlink:role="http://www.xbrl.org/2003/role/label"	rdfs:comment

```
@prefix csOnt: <http://lmad.ifs.tuwien.ac.at/CorporateSustainability#>.
   csOnt:CorruptionAspectTitle a owl:Class;
   csOnt:hasReferenceName csOnt:SustainabilityReportingGuidelines;
   csOnt:hasReferenceNumber csOnt:G3.1;
   csOnt:hasReferencePublisher csOnt:GlobalReportingInitiative;
   csOnt:hasReferenceSection csOnt:Profile;
   rdfs:comment "Corruption aspect [abstract]";
   rdfs:subClassOf csOnt:SocietyPerformanceIndicatorsTitle.
```

Listing 1: Transformation result represented in N3 notation due to better readability (own representation).

the general layout extracted from the GRI-taxonomy, the ontology can be extended as required by input sources or application components.

3.3 Mapping of Sustainability Data to the Ontology

As a next step, we illustrate how input sources can be mapped to the ontology to enable an improved semantic understanding. An exmemplary input source are statistical indicators provided by the World Bank offered as LOD through a third party[1]. Listing 2 illustrates how this data input may be semantically linked to the new ontology. The proposed SPARQL query retrieves the "Control of corruption" indicator of the World Bank LOD set for each country and each year and returns the result set in form of RDF triples (based on the CONSTRUCT operator). The statistical level of corruption in each country can be a relevant input for location-based risk analysis and is also included in the GRI-framework. Every datapoint is modeled as a different entity and directly linked to the GRI ontology using the property "dc:Subject" of the Dublin Core standard[2]. This property specifies the subject of a resource and, in this case, the respective GRI subject of the datapoint (see line 8 in Listing 2). In further research, the datapoints could be grouped and the respective indicator set specified in more detail.

Listing 3 shows the output of aforementioned SPARQL query in N3 notation. As expected, the resulting entities are linked, amongst others, to the automatically generated Corporate Sustainability ontology "csOnt". The returned datapoints have URIs (`<http://worldbank.270a.info/dataset/world-bank-indicators/GV.CONT.CO.ES/MZ/2007>`), are grouped by country (sdmx-concept:refArea) and year (sdmx-dimension:refPeriod) and hold the actual value as a literal (sdmx-measure:obsValue). This output may be used for further processing in the application, e.g. transferred as messages in the final system.

[1] `http://worldbank.270a.info/about.html`

[2] `http://dublincore.org/documents/dcmi-terms/#elements-subject`

4. EVALUATION

Since the ontology is based on a well-established standard in the business environment, it should adequately represent the business conceptualization of Corporate Sustainability (*Requirement 1*). Exploiting existing Xlink-information in the XBRL-Taxonomy, the hierarchical relations can be preserved and a basic structure is provided (*Requirement 2*). The exemplary illustration shows how an input source can be mapped to this ontology. However, Requirement 3 remains to be evaluated in future research, where a broader range of input sources needs to be mapped to the ontology. Summing up, the ontology provides a first basic representation of Corporate Sustainability. In a further step, an extension of this ontology using bottom-up information could be useful, e.g. by augmenting the ontology with concepts extracted from sustainability reports. This can be seen as a combination of a top-down with a bottom-up approach for ontology engineering that has also been already proposed in related domains (e.g. Hare et al. [13]).

5. RELATED WORK

Several attempts have been made to create ontologies in the domain of sustainability. Kumazawa et al. [16] suggest an ontological approach for the structuring of the concepts and relations within the field of sustainability science. Han and Stoffel [12] advocate for an ontological approach for the integration of qualitative case studies in the domain of environmental sustainability research. The approach of Pinheiro et al. [20] proposes an ontological approach that adaptively allows to relate sustainability indicators. Bertin et al. [2] present an ontology-based approach for the inventory database of life cycle assessments. Edum-Fotwe and Price [6] suggest a social ontology relating and structuring environmental, social and economic issues in order to achieve overall social sustainability. Giovanni et al. [9] suggest a product-centric ontology together with a system that allows sustainability-improved alternatives in manufacturing.

As can be seen, these approaches aim for the use of ontologies in the broader sustainability domain, however, none

```
PREFIX csOnt: <http://lmad.ifs.tuwien.ac.at/CorporateSustainability/>

CONSTRUCT {
        ?datapoint
                sdmx-concept:refArea ?country ;
                sdmx-dimension:refPeriod ?year ;
                sdmx-measure:obsValue ?value ;
                dc:subject csOnt:CorruptionAspectTitle .
} WHERE {
        ?datapoint property:indicator indicator:GV.CONT.CO.ES .
        ?datapoint sdmx-dimension:refArea ?country .
        ?datapoint sdmx-measure:obsValue ?value .
        ?datapoint sdmx-dimension:refPeriod ?year .
}
```

Listing 2: SPARQL CONSTRUCT query to retrieve "Control of corruption" indicators from the World Bank LOD repository. The result will be linked to the GRI sustainability ontology (own representation).

```
<http://worldbank.270a.info/dataset/world-bank-indicators/GV.CONT.CO.ES/MZ/2007>
 dc:subject   csOnt:CorruptionAspectTitle;
 sdmx-concept:refArea   country:MZ;
 sdmx-dimension:refPeriod   <http://reference.data.gov.uk/id/year/2007>;
 sdmx-measure:obsValue   -0.5.
```

Listing 3: Extract of the result of the SPARQL CONSTRUCT query from Listing 2 in N3 notation (own representation).

of these ontology specifically relates to the area of Corporate Sustainability and integrates the well- established GRI-standard.

Furthermore, several studies presented mechanisms to transform XBRL-taxonomies into ontologies in the financial domain [17, 7, 3, 1, 4, 22]. Others, focus on semantic retrieval of XBRL data [14] or suggest a semantic augmentation of the existing XBRL-standard [25, 19]. Ghahremanloo et al. [8] even propose an ontology for sustainability indicators integrating OECD and GRI frameworks. However, none of the existing approaches makes use of the full range of information available in the XBRL taxonomy, like hierarchical relations.

6. CONCLUSION & OUTLOOK

In this paper, we present the idea to utilize the GRI XBRL-taxonomy to build a basic Corporate Sustainability ontology. We show that XBRL-elements and XLinks can be transformed into an OWL class hierarchy and how actual sustainability data can be linked to this ontology, showcasing the applicability of our approach.

The resulting ontology can be used to overcome source heterogeneity in Corporate Sustainability information systems, as described in [24]. To the best of our knowledge, this is the first ontology in the domain of Corporate Sustainability emphasizing on the exploitation of its hierarchical relations based on the GRI standard.

Enriched with bottom-up information we have the idea to evaluate the ontology by including it into information extraction tasks that automatically identify Corporate Sustainability-relevant news articles or user-generated content (social media, SMS).

A further interesting field for future research is given by the possibility to not only transform taxonomies into ontologies, but also to transform Sustainability Reports represented in XBRL instance documents (XML) into RDF documents using these ontologies. Thereby, these data could be made queryable with SPARQL unleashing multiple opportunities to compare, link and visualize sustainability data from different corporations in an open manner.

7. ACKNOWLEDGMENTS

Lisa Madlberger is financially supported by the Vienna PhD School of Informatics. Peter Wetz is financially supported by the Doctoral College Environmental Informatics.

8. REFERENCES

[1] J. Bao, G. Rong, X. Li, and L. Ding. Representing financial reports on the semantic web:. In M. Dean, J. Hall, A. Rotolo, and S. Tabet, editors, *Semantic Web Rules*, number 6403 in Lecture Notes in Computer Science, pages 144–152. Springer Berlin Heidelberg, Jan. 2010.

[2] B. Bertin, V.-M. Scuturici, E. Risler, and J.-M. Pinon. A semantic approach to life cycle assessment applied on energy environmental impact data management. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pages 87–94, New York, NY, USA, 2012. ACM.

[3] H. Carretié. Improving the comparability of financial statements in a changeable context: a semantic web-based approach. 2012. Tesis Doctoral leída en la Universidad Rey Juan Carlos de Madrid en 2012. Directores de la Tesis: Ms. Beatriz Torvisco y Mr. Roberto García.

[4] T. Declerck and H.-u. Krieger. Translating XBRL into description logic: an approach using protege, sesame and OWL. In *In Proceedings of Business Information Systems (BIS)*, 2006.

[5] Z. Ding and Y. Peng. A probabilistic extension to ontology language OWL. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii international conference on*, page 10 pp, 2004.

[6] F. T. Edum-Fotwe and A. D. Price. A social ontology for appraising sustainability of construction projects and developments. *International Journal of Project Management*, 27(4):313–322, 2009.

[7] R. García and R. Gil. Triplificating and linking XBRL financial data. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 3:1–3:8, New York, NY, USA, 2010. ACM.

[8] L. Ghahremanloo, J. A. Thom, and L. Magee. An ontology derived from heterogeneous sustainability indicator set documents. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, pages 72–79, 2012.

[9] A. Giovannini, A. Aubry, H. Panetto, M. Dassisti, and H. El Haouzi. Ontology-based system for supporting manufacturing sustainability. *Annual Reviews in Control*, 36(2):309–317, 2012.

[10] Global Reporting Initiative. G3.1 sustainability reporting guidelines. Technical report, Global Reporting Initiative, Amsterdam, 2011.

[11] P. Gluchowski, R. Gabriel, and P. Chamoni. *Management Support Systeme und Business Intelligence: Computergestützte Informationssysteme für Fach-und Führungskräfte*. Springer-Verlag New York, Inc., 2005.

[12] D. Han and K. Stoffel. Ontology based qualitative case studies for sustainability research. In *Proceedings of the AI for an Intelligent Planet*, page 6, 2011.

[13] J. S. Hare, P. A. S. Sinclair, P. H. Lewis, K. Martinez, P. G. B. Enser, and C. J. S. Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches. In *In 3rd European Semantic Web Conference (ESWC-06), LNCS 4011*. Springer Verlag, 2006.

[14] M. Huang, D. Wang, and K. Wang. Ontology-based semantic retrieval of XBRL data. In *2011 International Conference on Business Computing and Global Informatization (BCGIN)*, pages 363–366, 2011.

[15] H.-G. Kemper, W. Mehanna, and C. Unger. *Business Intelligence - Grundlagen und praktische Anwendungen: Eine Einfhrung in die IT-basierte Managementuntersttzung: Eine Einführung in die IT-basierte Managementunterstützung*. Vieweg+Teubner Verlag, 2., erg. aufl. 2006 edition, Nov. 2006.

[16] T. Kumazawa, O. Saito, K. Kozaki, T. Matsui, and R. Mizoguchi. Toward knowledge structuring of sustainability science based on ontology engineering. *Sustainability Science*, 4(1):99–116, Feb. 2009.

[17] R. Lara, I. Cantador, and P. Castells. XBRL taxonomies and OWL ontologies for investment funds. In *Advances in Conceptual Modeling-Theory and Practice*, pages 271–280. Springer, 2006.

[18] U. Leser and F. Naumann. *Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen*. Dpunkt Verlag, 1. a. edition, 2007.

[19] B. Li and L. Min. An ontology-augmented XBRL extended model for financial information analysis. In *IEEE International Conference on Intelligent Computing and Intelligent Systems, 2009. ICIS 2009*, volume 3, pages 99–103, 2009.

[20] W. A. Pinheiro, R. Barros, J. M. D. Souza, G. B. Xexeo, J. C. M. Strauch, P. M. D. Barros, and M. P. Campos. Adaptative methodology of sustainability indicators management by ontology. *International Journal of Global Environmental Issues*, 9(4):338–355, 2009.

[21] M. E. Porter and M. R. Kramer. The link between competitive advantage and corporate social responsibility. *Harvard Business Review*, 11, 2006.

[22] M. Spies. An ontology modelling perspective on business reporting. *Inf. Syst.*, 35(4):404–416, June 2010.

[23] R. Stevens, C. A. Goble, and S. Bechhofer. Ontology-based knowledge representation for bioinformatics. *Briefings in bioinformatics*, 1(4):398–414, 2000.

[24] A. Thöni, L. Madlberger, and A. Schatten. Towards a data-integration approach for enterprise sustainability risk information systems. In J. Basel, P. Jasek, O. Novotny, and A. M. Tjoa, editors, *Proceedings of the 7th International Conference on Research and Practical Issues of Enterprise Information Systems*, Linz, 2013. Trauner Verlag.

[25] H. Zhu and S. E. Madnick. Semantic integration approach to efficient business data supply chain: Integration approach to interoperable XBRL. Working paper, Jan. 2008.