Knowledge Organisation for Digital Libraries

Lena-Luise Stahn[1], Ingetraut Dahlberg, and Ernesto William De Luca[1]

Georg Eckert Institute Leibniz Institute for International Textbook Research
stahn@leibniz-gei.de, Ingetraut.Dahlberg@t-online.de,
deluca@leibniz-gei.de

Abstract. Today research data and output from nearly every disciplinary and interdisciplinary domain exist also (and sometimes only) in digital form. Consequently, to ensure the data's efficient retrieval as well as long term usage and relevance, Knowledge Organisation and its tools/systems, i.e. lexical resources like thesauri and ontologies, play a major role in the Digital Libraries world.
In this paper we discuss our approach of pursuing the conversion and mapping of two of the most promising lexical resources, the Information Coding Classification and the MultiWordNet, after having converted them into the EuroWordNet RDF/OWL format. In a second step we show how to integrate this additional knowledge into a domain Knowledge Organisation System (KOS). In the end we will have a presentation of the method and mapping as well as a use case for its evaluation for Information Retrieval purposes.

Keywords: Knowledge Organisation, Lexical Linked Open Data, Digital Libraries

1 Introduction

1.1 Knowledge Organisation Systems (KOS), Linked Open Data (LOD) and Digital Libraries: Current Situation and Problems

The relevance of Knowledge Organisation has gained in importance with the emergence of the digital era and the World Wide Web. Just as in the physical world these large arising digital libraries require adequate and thorough structuring systems in order to organise the increasing amount of available data and to avoid loss of information. Making them accessible by "putting them into the Semantic Web" requires thorough description and homogenisation, i.e. the provision of a data model capable of dealing with large amounts of data from various domains.

"A KOS serves as a bridge between the user's information need and the material in the collection. With it, the user should be able to identify an object of interest without prior knowledge of its existence."[1]

[1] https://www.clir.org/pubs/reports/pub91/1knowledge.html

The Georg Eckert Institute[2] (GEI) deals with various Textbook Research resources from multiple disciplines. Research data, produced within in a specific context under particular research questions, results a broad spectrum of digital resources, e.g. a digital collection of historic textbooks[3], repositories for curricula[4] and textbook research related publications[5], as well as digital information services[6].

Because there still does not exist a proper Knowledge Management at the GEI as would be provided by using a suitable KOS these research results can not yet be linked to other research contexts and therefore are in danger of not being used afterwards. Their production within a highly specialized research community with complex but separate contexts and systems prevents them from being found easily, consequently followed by death of data", double work and waste of resources. Another researcher with related information interests has no knowledge about the already existing GEI data and is not able to satisfy their need of information easily.

Although many data models and classification systems have been developed to help here, their inadequacy is evident, as these models, too, have their intrinsic structures and characteristics and are most often bound to their respective domain. Digital Libraries suffer from too specialised models, which are not able to communicate with each other. The advantages of the Linked Data World (Berners-Lee et al. 2001) are therefore not being exploited (De Luca/Dahlberg 2014, 9). Our idea is to transport this knowledge into the Digital Libraries world. With this a link may be enabled, e.g. between two disciplines seemingly as slightly related as the textbook research and the restoration domain, which eventually enriches both knowledge domains. Pursuing our preliminary work (De Luca/Dahlberg 2014) in section 2.2, we will show our approach of improving this situation by using broader KOS like Top Ontologies and Lexical Resources (e.g. WordNet Domains and Information Coding Classification) and how to integrate these into the Linked Data Cloud. In this paper, we give an overview of Ingetraut Dahlberg's Information Coding Classification and its potential for Information Retrieval and Knowledge Organisation in the Digital Libraries world (see section 2.2.1). Section 2.2.2 covers Ernesto William De Luca's work of converting EuroWordNet into the Semantic Web-compliant format RDF/OWL. Both their procedures will be adapted in this work, as shown in section 3.1 and 3.2, followed by a proof of concept (3.3). Because of additional expertise in the field, our current approach intends to pursue this work by applying the restoration domain as a use case: if the domain knowledge's integration into the (Multilingual) Lexical Linked Data Cloud succeeds, it will be enriched with knowledge from a broader KOS as is provided by the ICC with its Knowledge Domains. Consequently a link will be created other research fields for which its relevance has not yet been

[2] http://www.gei.de/en/home.html
[3] http://gei-digital.gei.de/viewer/
[4] http://bibliothek.gei.de/en/library/curricula-workstation-textbookcat.html
[5] http://repository.gei.de/
[6] http://edumeres.net/

detected. Ideally new research questions may arise from this unexpected combination. Especially the field of textbook research with its highly multidisciplinary character may benefit (cf. Dahlberg 2017, 2014, 120).

1.2 Preliminary Work

Dahlberg's Information Coding Classification Dahlberg's Information Coding Classification (ICC) (Dahlberg 2017, 2014, 1974) system follows this intention with its structure based on a model of ordering the world's knowledge disciplines according to "evolution theoretic aspects" ("evolutionstheoretischen Gesichtspunkten"). The knowledge of the world is seen as belonging to different levels of existence ("Seinsschichten"), each structured through "aspects". Structuring each (sub-)level this aspect-oriented view features infinite extensibility. Its bilingual character can be seen as an additional advantage: The multilingualism of the Web is a source of ambiguities where the ICC may function as a translation tool.

The initial development of the ICC as a "faceted classification" by Dahlberg in the early seventies was made possible through a funding by the former Dt. Gesellschaft fr Dokumentation e.V. (DGD)[7]. Dahlberg's intention was to establish a classification system not depending on the division into knowledge disciplines in order to facilitate long term usage and extensibility (De Luca/Dahlberg 2014, 3 et seq.). Her approach is based on J.K.Feibleman's (1954) and N. Hartmann's (1964) "Schichtentheorie des Seins" ("Theory of Levels of Being"), dividing existence/being into layers of "areas of being" ("Seinsbereiche"). Each layer is structured by nine "categorically defined aspects" functioning as "Systemstellenplan" oder "Systemifikator" and dividing the "areas of being" into nine "Subject Groups" ("Sachgruppen") (SG) and further into nine "Knowledge Domains" ("Wissensgebiete") (WG) (for further description cf. Dahlberg 2017, 2014, 1974, 230 et seq., 259 et seq.).

De Luca's RDF/OWL for EuroWordNet De Luca (De Luca/Dahlberg 2014, 4) pursued the findings in the context of improving Information Retrieval results by means of NLP techniques. He decided to take the ICC into consideration in his approach of extending the classification resources WordNet Domains (Magnini/Cavagli 2000) (WN Domains),based on the commonly used Princeton WordNet (Miller et al. 1990, Fellbaum 1998) (PWN), and especially the multilingual extension EuroWordNet (Vossen 1997) (EWN), used in the Semantic Web context. De Luca's work comprises the conversion of EuroWordNet into a Semantic Web-compliant format (De Luca et al. 2007), for which RDF/OWL has proved to be the most useful, based on the RDF/OWL format for Princeton WordNet developed in (van Assem et al. 2004). The RDF/OWL WordNet model comprises of three main classes SynSet, WordSense and Word with respective subclasses, and three relation types (for SynSet and WordSense each,

[7] In 2014 it was renamed as Deutsche Gesellschaft fr Information und Wissen e. V. (DGI). http://dgi-info.de/

0 GENERAL FORM CONCEPTS	01 THEORIES; PRINCIPLES	02 OBJECTS; COMPONENTS	03 ACTIVITIES PROCESSES	04 PROPERTIES or 1st kind of field specialty	05 PERSONS or 2nd kind of Field specialty	06 INSTITUTIONS or 3rd kind of field specialty	07 TECHNOLOGY & PRODUCTION	08 APPLICATION in other fields, DETERMINATION	09 DISTRIBUTION & SYNTHESIS
1 FORM & STRUCTURE AREA	11 Logic	12 Mathematics	13 Statistics	14 Systemology	15 Organization Science	16 Metrology	17 Cybernetics, Control & Automation	18 Standardization	19 Testing & Monitoring
2 MATTER & ENERGY AREA	21 Mechanics	22 Physics of Matter	23 General & Technical Physics	24 Electronics	25 Physical Chemistry	26 Pure Chemistry	27 Chemical Technol. & Engineering	28 Energy Science & Technology	29 Electrical Engineering
3 COSMOS & EARTH AREA	31 Astronomy & Astrophysics	32 Astronautics & Space Research	33 Basic Geosciences	34 Atmospheric Sciences & Technology	35 Hydrospheric & Ocean. Science & Technology	36 Geological Sciences	37 Mining	38 Materials Science & Technology	39 Geography
4 BIO SPHERE	41 Basic Biological Sciences	42 Microbiology & Cultivation	43 Plant Biology & Cultivation	44 Animal Biology & Breeding	45 Veterinary Sciences	46 Agriculture & Horticulture	47 Forestry & Wood Science & Technology	48 Food Sciences & Technology	49 Ecology & Environment
5 HUMAN AREA	51 Human Biology	52 Health & Theor. Medicine	53 Pathology & Pract. Medicine	54 Clinical Medicine & Cure	55 Psychology	56 Education	57 Occupation, Labor &Leisure	58 Sports	59 Household & Home Life
6 SOCIETAL AREA	61 Sociology	62 State & Politics	63 Public Administration	64 Money & Finances	65 Social Assistance, Appraisal & Survey	66 Law & Legal Science	67 Areal Planification & Urbanism	68 Structure of Defense	69 History Science & History
7 ECONOMY & TECHNOLOGY AREA	71 General & National Economics	72 Applied Economics, Business Mgt.	73 Technical Sciences	74 Mechanical & Precision Engg.	75 Building & Civil Engineering	76 Science of Commodities & Technol.ofGoods	77 Vehicle Science & Technology	78 Traffic & Transport.Techn. & Services	79 Service Economics
8 SCIENCE & INFORMATION AREA	81 Science of Science	82 Information Sciences	83 Computer Science	84 Information in general	85 Communication	86 Mass Communication	87 Printing & Publishing	88 Tele-communication	89 Semiotics
9 CULTURE AREA	91 Language & Linguistics	92 Literature & Philology	93 Music & Musicology	94 Fine Arts	95 Theatre	96 Culture Science (narrow sense)	97 Philosophy	98 Religion (in general)	99 Christian Religion & Theology

Fig. 1. Table 1: Information Coding Classification. Survey of Subject Groups (English)

and attributes). An example is shown in figure 1. The conversion supports standardisation and hence the long term usage, as well as the EuroWordNet's multilingualism. In (De Luca/Dahlberg 2014) we have adopted their EuroWordNet conversion for the ICC and discussed our approach how to extend the ICC with the EuroWordNet in RDF/OWL. This approach will be followed in the present paper (see section 3.2), using the MultiWordNet (Magnini et al. 1994), a multilingual version of the Princeton WordNet, which, as opposed to the EWN, also adheres to the PWN's structure and hence to the WN Domains. This way we will be able to implement the mapping between the WN Domains and the ICC Knowledge Domains, which at present has been generated on a theoretical level.

```
<ewn20schema:NounSynset rdf:about="&ewn20instances;synset-bank-noun-1"
    rdfs:label="bank">
    <ewn20schema:synsetId>102690337</ewn20schema:synsetId>
</ewn20schema:NounSynset>
<ewn20schema:Word rdf:about="&ewn20instances;word-bank"
    ewn20schema:lexicalForm="bank"/>
<ewn20schema:NounWordSense rdf:about="&ewn20instances;wordsense-bank-noun-1"
    rdfs:label="bank">
    <ewn20schema:word rdf:resource="&ewn20instances;word-bank"/>
</ewn20schema:NounWordSense>
<rdf:Description rdf:about="&ewn20instances;synset-bank-noun-1">
    <ewn20schema:containsWordSense rdf:resource="&ewn20instances;wordsense-bank-noun-1"/>
    <ewn20schema:containsWordSense rdf:resource="&ewn20instances;wordsense-bank_building-noun-1"/>
</rdf:Description>
```

Fig. 2. Figure 1: RDF/OWL-EuroWordNet SynSet Example

2 Approach extension and Use Case

2.1 ICC Actualisation

The work's continuation presented in this paper comprises several steps: the first step will be the actualisation and completion of the ICC itself, namely the integration of the fourth to sixth level. This will be achieved by completing Dahlberg's preliminary project "Logstruktur" (Dahlberg 1977, 1979) and pursuing the work of her dissertation (Dahlberg 1974). The following tasks need to be handled:

1. Revision: Sighting of the work done in 1974 by Dahlberg and complementing possibly missing definitions in all divisions, English translation.
2. Actualisation/Updating: inserting newly established knowledge domains into ICC (since the "Vademecum Deutscher Lehr- und Forschungsstten" (1960) used by Dahlberg for this step in 1974 does not exist anymore, a new source signing German research institutions needs to be found; domain experts' questioning).
3. Review by domain experts
4. compilation of defined terms in alphabetical order (Lexicon basis)
5. compilation of defined terms in systematic order (classification basis)
6. English translation

2.2 ICC and MultiWordNet conversion into RDF/OWL EWN and mapping

This part applies the conversion method from (De Luca et al. 2007) and (De Luca/Dahlberg 2014, 8), first format adaptation:

1. schema extension/adaptation:
   - ICC and MultiWordNet analysis: specific requirements and format adaptation
   - RDF/OWL EWN format adaptation for ICC and MultiWordNet
2. The second step will be to apply the conversion method developed in (De Luca et al. 2007) and again proposed in (De Luca/Dahlberg 2014) in order to have the ICC and MultiWordNet data in an adequate format which allows the theoretical and practical mapping onto the WN Domains:
   - theoretical mapping: map the ICC knowledge domains onto the WN Domains, based on the initial approach done in (De Luca/Dahlberg 2014, 6), taking into account the ICC updates from step one
   - practical mapping: After the two ontologies have been mapped on a theoretical as well as on a schema level, the actual data integration between the ICC and the WordNet Domains via the MultiWordNet will be possible. Again this procedure will be done according to the approach in (De Luca/Dahlberg 2014, 8 et seq.): enlarging the MultiWordNet coverage with the ICC generic terms of the Knowledge Domains, declaring every ICC generic term as a class (owl:Class) and every underlying term as a

subclass (rdfs:subClassOf), eventually supplementing every class (of the ICC top ontology) with a language description (e.g. xml:lang="en"), in order to add it to the correct set of language files of the MultiWordNet.

2.3 Proof of concept: Restoration Domain as Use Case and Information Retrieval Tasks

As a proof of concept, we decided to add a third step to link the restoration domain ontology intended for document indexing purposes: the chosen restoration terminologies have been established especially for the SemRes project and used only on an experimental level.

The integration is based on the approach in (De Luca et al. 2007) and will be done on a manual level: after determining the correct SynSet in the WN-ICC-Extension, the domain ontology's top concept will be added as a hyponym of the SynSet[8]. For this converting the domain ontologies into the same RDF/OWL format is necessary[9].

Eventually we will be able to perform Information Retrieval tasks with the converted KOS, which will test the newly integrated ICC top ontology for Information Retrieval purposes. We imagine use cases, in which Information Retrieval is done in the restoration domain, searching for the semantic concept "democracy". Through the newly established links between restoration domain ontology and the ICC Knowledge Domains via MultiWordNet, information on the "democracy" concept may also be found in the textbook research domain, where the concept would be used in the context of education or art history. This way, abstraction of concepts, integrated into the available Lexical Linked Data Cloud, will be facilitated, and the discovery of relations between different domains will be allowed.

3 Conclusion and future work

In this paper we have shown our approach how to establish a "Lexikon der Wissensgebiete" by pursuing the work of Dahlberg (1974, 1977, 1979). We then discussed how to bring this knowledge database into the lexical linked data cloud by converting it into RDF/OWL to use it as a Top Ontology within the Lexical Resource WordNet. This step is based on the EuroWordNet conversion into RDF/OWL, presented by De Luca et al. (2007). By integrating the restoration

[8] "After having found the correct SynSet, we merged manually the complete converted domain ontology under the appropriate hyperonym (SynSet). In this case we could enlarge the EuroWordNet coverage with domain-specific terms." De Luca et al. 2007, 9.

[9] "The first step before including the domain ontologies in the new EuroWordNet OWL hierarchy was to convert these in the same OWL format [] merging methods for including these domain-ontologies to the EuroWordNet Owl representation [] domain ontology is then added to the generic one, directly under its new hyperonym." De Luca et al. 2007, 9.

domain ontology as a use case we can evaluate the usefulness of the enriched lexical resource for Information Retrieval or Query expansion tasks.

Further possible work comprises the connection of additional domain ontologies, e.g. archaeology, to evaluate the usefulness of the produced and enriched lexical linked data knowledge repository in Digital Humanities research scenarios. We imagine this to offer a general improvement for Information Retrieval within the Digital Libraries world.

References

1. van Assem M., Gangemi A., and Schreiber G.: Wordnet in RDFS and OWL. Technical report, W3C (2004)
2. Berners-Lee, T., Hendler, J., and Lassila, O.: The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, 284(5), 1-5 (2001)
3. Dahlberg, I.: Grundlagen universaler Wissensordnung. (1974)
4. Dahlberg, I.: Logstruktur Ein Projekt zur Generierung einer Systematik der Wissenschaften. Deutsche Universittszeitung 8, 250 (1977)
5. Dahlberg, I.: Das DFG-Projekt Logstruktur". Formaler Abschlussbericht. (1979)
6. Dahlberg, I.: Information Coding Classification: Geschichtliches, Prinzipien, Inhaltliches. Information Wissenschaft und Praxis 61(8), 449-454 (2010)
7. Dahlberg, I.: A faceted classification of general concepts. Classification & Ontology. Proc.Intern. UDC-Seminar 2011, 177-193 (2011)
8. Dahlberg, I.: A systematic new lexicon of all knowledge fields, based on the Information Coding Classification. Knowledge Organization 39(2), 142-150 (2012)
9. Dahlberg, I.: Wissensorganisation: Entwicklung, Aufgabe, Anwendung, Zukunft. Textbooks for Knowledge Organization, Vol 3 (2014)
10. Dahlberg, I.: Brief Communication: Why a new universal classification system is needed. Knowledge Organization, 44(1), 65-71 (2017).
11. De Luca E. W., Eul M., and Nrnberger A.: Converting EuroWordNet in OWL and Extending It with Domain Ontologies. In: Proceedings of the Workshop on Lexical-Semantic and Ontological Resources, ## (2007)
12. De Luca E. W.: Semantic Support in Multilingual Text Retrieval. (2008)
13. De Luca, E. W., Plumbaum, T., Kunegis, J., and Albayrak, S.: Multilingual Ontology-based User Profile Enrichment. In: Proceedings of the First International Workshop on the Multilingual Semantic Web (MSW 2010), in conjunction with WWW 2010 - 19th International World Wide Web Conference 571, 41-42 (2010)
14. De Luca, E. W.: Extending the Linked Data Cloud with Multilingual Lexical Linked Data. Knowledge Organization, 40(5) (2013)
15. De Luca, E. W., and Dahlberg, I.: Die Multilingual Lexical Linked Data Cloud: Eine mgliche Zugangsoptimierung? Information-Wissenschaft & Praxis, 65(4-5), 279-287 (2014)
16. Feibleman, J.K.: The Integrative Levels in Nature. British J. Philosophy of Science, May (1954)
17. Fellbaum, C.: WordNet, an electronic lexical database. (1998)
18. Hartmann, N.: Der Aufbau der realen Welt. Grundriss einer allgemeinen Kategorienlehre. (1964)

19. Magnini, B., Strapparava, C., Ciravegna, F., and Pianta, E.: Multilingual lexical knowledge bases: Applied WordNet prospects. In: Proceedings of the International Workshop on "The Future of the Dictionary" (1994)
20. Magnini, B., and Cavagli, G.: Integrating Subject Field Codes into WordNet. In: LREC, 1413-1418 (2000)
21. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J.: Introduction to WordNet: An online lexical database. International journal of lexicography, 3(4), 235-244 (1990)
22. VDLF ; ein Handbuch des wissenschaftlichen Lebens. (1960)
23. Vossen, P.: EuroWordNet: a multilingual database for information retrieval. In: Proceedings of the DELOS workshop on Cross-language Information Retrieval, 5-7 (1997)