Linked Thesauri Quality Assessment and Documentation for Big Data Discovery

Riccardo Albertoni, Monica De Martino, Alfonso Quarati
Institute for Applied Mathematics and Information Technologies "E. Magenes"
National Research Council of Italy, Genoa
{albertoni,demartino,quarati}@ge.imati.cnr.it

***Abstract*—Thesauri are knowledge systems which may ease Big Data access, fostering their integration and re-use. Currently several Linked Data thesauri covering multi-disciplines are available. They provide a semantic foundation to effectively support cross-organization and cross-disciplinary management and usage of Big Data. Thesauri effectiveness is affected by their quality. Diverse quality measures are available taking into account different facets. However, an overall measure is needed to compare several thesauri and to identify those more qualified for a proper reuse. In this paper, we propose a Multi Criteria Decision Making based methodology for the documentation of the quality assessment of linked thesauri as a whole. We present a proof of concept of the Analytic Hierarchy Process adoption to the set of Linked Data thesauri for the Environment deployed in LusTRE. We discuss the step-by-step practice to document the overall quality measurements, generated by the quality assessment, with the W3C promoted Data Quality Vocabulary.**

Keywords: quality; linked data; thesauri; AHP; metadata; DQV

I. INTRODUCTION

The 3Vs (Volume, Variety and Velocity) Big Data model originally envisaged by Doug Laney in his 2001 seminal report [1], constantly evolved in the past years to include more Vs, from the seven ones (i.e. Variability, Veracity, Visualization, and Value) summarized by Eileen McNulty[1], up to the ten suggested by Kirk Borne[2] who added three further Vs (Venue, Vocabulary and Vagueness). This V-based characterization of Big Data serves to highlight its major challenges: the acquisition, cleaning, curation, validation, integration, storage, processing, indexing, search, analysis of huge volumes of relentless multifaceted data. Underlying these issues is the need of weaving together the myriad connections scattered from such multitude of information into a cohesive network capturing how every piece of data fits together into the global picture [2]. To this end, different Big Data sources may be organized in ways that distinct concepts with similar meanings may be connected by semantic metadata [3].

This is important for, often not structured and neatly formatted, information produced by Web applications. Data including text, images, video, and audio formats, in order to be discovered and properly processed by most demanding applications, require to be stored, queried, and integrated across this variety of information types [4]. Linked Data may serve this purpose, by providing an architectural pattern for mapping, connecting and indexing heterogeneous information from different sources [5]. They allow to represent information in human or machine-readable form, fostering new relationships to be inferred from existing data. They facilitate the analysis and searching of Big Data systems [6]. As publishing paradigm Linked Data enable the extension of the Web into a global data space based on open standards, the so-called Web of Data [7]. According to Bizer et al., "the Web of Data covers a wide variety of topics ranging from data describing people, organizations and events, over products and reviews to statistical data provided by governments as well as research data from various scientific disciplines." [8] This variety also applies to other dimensions, such as the representation of formats, data models, basic conceptualizations, correctness, etc. [9].

Almost 90% of Big Data tends to be unstructured[3] (i.e. has no formal schema) thus metadata (data about data) become important. Linked Data provide a view of the bigger picture by tying together heterogeneous records [6]. Data on the Web will not be discoverable or reusable if insufficient metadata is provided [10]. Linked Data may support publishers in describing data accurately through comprehensive metadata of Big datasets. These metadata provide relevant information about authorship, currency, licensing terms, which help data consumers in the discovery and reuse of datasets [7] [11]. To represents the variety of common types and entities collected in Big datasets, Linked Data sources reuse terms from widely-used vocabularies making it easier for applications to mine and understand data from different data repositories [8].

Terminology sources like thesauri, taxonomies and controlled vocabularies support indexing, organization and search of both structured and unstructured information. The Simple Knowledge Organization System[4] (SKOS) standard is promoted by the World Wide Web Consortium to supply a common data model for sharing and linking thesauri on the Web. SKOS thesauri are published according to the Linked Data principles and expressed as Resource Description

[1] http://dataconomy.com/seven-vs-big-data/
[2] https://www.mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs
[3] https://vitalrecord.tamhsc.edu/big-data-health-care-revolution-7-vs-big-data/
[4] http://www.w3.org/TR/skos-reference

Framework[5] (RDF) triples. SKOS-based Linked Data thesauri are used to effectively manage Big Data. SKOS allows the connection of multiple thesauri in order to create cross-browsing and cross-searching applications for Big Data [12].

Through a combination of Semantic Web standards like RDF, SKOS and OWL[6], Linked Data might ease the access to big datasets fostering their integration and re-use. In particular, Linked Data can help reducing Big Data variability: the adoption of RDF as basic data representation language for Linked Data may decrease several syntactic issues. However, as noted by Mitchell and Wilson [6] Linked Data "is no panacea – if rigorous controls are not applied to the meta model then it becomes yet another unstructured data source, making the problem worse, rather than better!". Furthermore, missing or incorrect metadata precludes consumers from finding relevant data. Even worse, if a user finds interesting datasets, but affected by outdated links, with data not compliant with the format declared in the metadata, or containing erroneous data, misleading links faulty syntax, broken links, conflicting, or intentionally wrong (e.g. spam) [8][11]. Such quality issues may inhibit or slow down the re-use of big datasets, discouraging potential users and undermining data publishers' efforts. Therefore, as recommended in [10] "the inclusion of data quality information in data publishing and consumption pipelines is of primary importance". Furthermore, ranking datasets, based on quality assessment by previous users, may foster their reuse [8] [13]. These considerations also hold true for the case of linked thesauri, used as a means to provide access to the Web of Data. Thus, the quality assessment of linked thesauri is particularly relevant for identifying those to be reused according to different contexts of use [14]. Several quality dimensions have been defined in the Linked Data field [15] and others more specific for thesauri described through the SKOS model [16]. A critical issue is to combine all these dimensions in order to compare thesauri, thus facilitating decision maker in choosing the ones better fitting her needs.

The paper presents a Multi Criteria Decision Making based methodology for the thesauri quality assessment. The proposed approach is aimed at supporting decision makers in thesauri comparison, through the exploitation of an overall quality measure. This measure takes into account the subjective perceptions of the decision maker according to her needs. The Analytic Hierarchy Process (AHP) methodology [17] is adopted to capture both subjective and objective facets involved in the thesauri quality assessment and to provide a ranking of the assessed thesauri. In [18] we provided a proof of concept of the AHP adoption to the set of linked thesauri within the thesaurus framework for the Environment LusTRE [19] developed within the EU project eENVplus[7]. Currently LusTRE's activity is focused on the assessment of the linked thesauri quality. This process bases on the evaluation of some SKOS quality criteria computed by the qSKOS tool[8] and results in the overall ranking of the assessed thesauri.

Herein, we also discuss the formalization of the thesauri quality assessment outcomes, according to the recommendation of the W3C Data on the Web Best Practices Working Group (DWBP-W3C) of "providing metadata as a fundamental requirement when publishing data on the Web". The documentation of the thesauri quality assessment process aims at making it intelligible and (possibly) replicable in other contexts of use. The publication of quality metadata is facilitated by the extension of the qSKOS tool [20] we have carried out. The extended qSKOS allows the automatic production of quality documents compliant with the metadata standard of the Data Catalog Vocabulary (DCAT), designed to foster interoperability between data catalogs published on the Web [21], and the W3C Data Quality Vocabulary (DQV) [22].

The paper is structured as follows: Section II presents the related works and a background of the DQV; Section III introduces the thesauri quality assessment and the AHP methodology; Section IV shows the adoption of AHP to LusTRE; Section V exemplifies how documenting the overall quality through the DQV; Section VI concludes.

II. Related Works And Background

A. Information Quality

The Information Quality (IQ) issue [23] was encountered in the Linked Data field of which SKOS thesauri are a particular case. Various proposals arose in these last years, addressing specific aspects of linked datasets quality and proposing specific sets of metrics and methodologies for their evaluation [24][25][26]. Zaveri et al. [15] provided a systematic survey on quality assessment for Linked Data: they identified 18 IQ dimensions, and classified them into four major categories. Looking at the problem of the quality of thesauri, Kless and Milton [27] suggested a range of abstract measurements based on quality notions in thesaurus literature. The declared purpose of these measurements is to support the evaluation approaches of thesauri but, they are solely based on theoretical analysis. The Authors themselves pointed out the necessity of operationalizing the measures and refined them by an application to real cases of thesauri. In [28] we extended [24] by proposing the "linkset importing" as a novel quality measure which estimates the completeness of dataset obtained by complementing SKOS thesauri with their skos:exactMatch related information. Such measure focused on easing multilingual issues such as incomplete language coverage, which affects many of the most popular SKOS thesauri. Suominen and Mader [16] provided one of the most complete related work about SKOS thesauri quality that introduces a set of 26 quality issues, defined as computable functions exposing potential quality problems. By using the qSKOS[9] tool they analyzed a corpus of 24 vocabularies, checking for their quality with respect to the 26 identified issues. Authors presented several facets of the vocabulary quality assessment and supplied useful recommendations and best practices.

From the related works analysis, we noticed that the most of the quality methodologies focus on single metrics rather than

[5] https://www.w3.org/RDF/
6 https://www.w3.org/OWL/
7 http://www.eenvplus.eu/
8 https://github.com/cmader/qSKOS/

[9] https://github.com/cmader/qSKOS

on how to aggregate and combine them. The aggregation of metrics was partially addressed in the framework Luzzu [26] and in the quality model proposed in [29]. Luzzu introduces a feature enabling users to allocate weights to their preferred categories, dimensions or metrics that are deemed suitable for their specific task. Based on these weights, ranks are obtained by using a simple weighted sum on all the metrics. The quality model proposed in [29] provides a unique terminology and reference for Linked Data quality specification and evaluation. The model specifies a set of quality measures related to Linked Data, together with formulas for their calculation, allowing measures aggregation with different levels of details.

B. *Metadata vocabularies*

Different initiatives stress the importance of metadata to access datasets as well as to document data quality. The DWBP-W3C Working Group [10] recommends the provision of machine-readable metadata, including quality information. The Research Data Alliance sets the Metadata Standards Directory Working Group, to support the development, implementation and use of metadata for scientific data[10].

Several metadata models were proposed so far. Dublin Core Metadata Initiative promoted DCTerms [11] metadata terms, which includes 15 core metadata elements popularized by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[12]. Upon the DCTerms, W3C promoted the DCAT vocabulary [21] and the Vocabulary of Interlinked Datasets (VoID) [13], two RDF vocabularies designed respectively to facilitate interoperability between data catalogs published on the Web and to express metadata about RDF datasets. The Coordination Group of the Interoperability Solutions for European Public Administrations (ISA) Programme defined the DCAT Application profile (DCAT-AP)[14] for describing public sector datasets in Europe. It is adopted as the common vocabulary for harmonizing descriptions of over 258,000 datasets harvested from 67 data portals of 34 countries. In the context of the development and maintenance of DBpedia [30], the metadata model DataID [31] combines and extends DCAT, VoID and DCTerms in order to explicitly deal with important aspects underspecified in the aforementioned vocabularies. A debate about the needs and challenges of harmonizing the emerging metadata models was recently facilitated in the W3C SDSvoc workshop[15]. An extent of interoperability among these models is granted by the re-use and the extension of a common set of metadata vocabularies (i.e., DCTerms, DCAT). However, brand new terms are often introduced to face with domain and community-specific requirements. New community efforts are under consideration, in particular to include new terms that can be relevant for a wider audience into DCAT.

Metadata describing data quality might be considered for DCAT extensions. Two vocabularies specifically designed to deal with quality documentation are the Dataset Quality Ontology[16] and the W3C DQV. The former provides a generic core vocabulary, allowing a uniform definition of specific data quality metrics. This definition would allow publishers to describe the quality benchmarks of their datasets. The DQV revises and extends the Dataset Quality Ontology to meet the requirements of the W3C DWBP working group. It builds as much as possible upon current vocabulary (e.g., DC, DCAT, SKOS, Prov-O[17], Web Annotation[18]) in order to maximize the reuse of standards and minimize the number of new terms introduced. The DQV deals with the quality of a wider set of kinds of data. It relies on the concepts of quality *metric*, *dimension* and *category*, widely discussed in [15]. Besides measurements from quality metrics, the DQV considers certificates, standards, and quality policies.

C. *The Data Quality Vocabulary (DQV)*

The DQV [22] supports the provision of quantitative or qualitative information about the dataset or its distributions. A quality metric (`dqv:Metric`) gives a procedure for measuring a data quality dimension, which is abstract, by observing a concrete quality indicator. Quality metric usually refers to quality dimensions (`dqv:Dimension`) which are quality-related characteristic of a dataset relevant to the consumer (e.g., the availability of a dataset). Dimensions, in turn, are grouped into categories (`dqv:Category`) in which a common type of information is used as a quality indicator. There are usually multiple metrics per dimension; e.g. the availability dimension can be indicated by the accessibility of a SPARQL endpoint, or that of an RDF dump or CVS file. Differently from the Dataset Quality Ontology, the DQV explicitly reuses SKOS in order to represent metrics, dimensions and categories. As a consequence, these concepts are described by means of lexical properties such as `skos:description`, and `skos:prefLabel`. Quality measurements (`dqv:QualityMeasurements`) are the outcomes produced by applying a certain metric. The actual gauged values are represented with the property `dqv:value`. They can be numeric (e.g., for the metric "human-readable labeling of classes, properties and entities", which measures the percentage of entities having an `rdfs:label` or `rdfs:comment`) or boolean (e.g., whether or not a SPARQL endpoint is accessible). The properties `dqv:computedOn` and `dc:date` represents respectively the dataset and the date on which the measurements is taken. `dqv:isMeasurementOf` is the property that associates each measurement to its metric. Information can be derived from other quality information. For example, measurements can be derived from other measurements. Metrics can be derived from other metrics. DQV models such derivations by means of the Prov-O ontology through the property `prov:wasDerivedFrom`.

III. QUALITY OF SKOS THESAURI

From the analysis of the related works we noticed that, apart from the framework Luzzu [26] and the quality model presented in [29], other proposals do not address the problem

[10] https://www.rd-alliance.org/groups/metadata-standards-directory-working-group.html
[11] http://dublincore.org/documents/dcmi-terms/
[12] http://www.openarchives.org/pmh/
[13] https://www.w3.org/TR/void/
[14] https://joinup.ec.europa.eu/system/files/project/dcat-ap_version_1.1.pdf
[15] https://www.w3.org/2016/11/sdsvoc/
[16] http://publica.fraunhofer.de/documents/N-351182.html
[17] https://www.w3.org/TR/2013/REC-prov-o-20130430/
[18] https://www.w3.org/TR/annotation-vocab/

of the quality of datasets as a whole. The two cited works, however, did not give major insights regarding the aggregation methodology, whereas in this paper we based on a well-founded decision-making technique. Our approach supports and documents a context-tailored aggregation of the various metrics and the overall quality assessment of linked datasets.

A. Thesauri quality assessment

In [14] we discussed how the well-founded decision-making technique AHP may support thesauri quality assessment and supply thesauri ranking. The ranking is obtained by synthetizing, for each thesaurus, an overall score computed from the aggregation of several IQ dimensions. Often, the assessment of IQ dimensions is made under an "objective" perspective, without considering the "subjective" point of view of the expert. Dealing with subjectivity allows asserting the importance of some dimension, or supply a judgment on dimensions that cannot quantitatively be measured (by a procedure) and required qualitatively assertion about their importance in a scenario. Indeed, IQ is strictly related to the context of use: for a given dataset, a variety of assessment values may be reported according to the aims, task and roles of the actors involved [32].

For what concerns the quality dimensions, our proposal relies on the work of Suominen and Mader [16] which provides a thorough discussion of quality *issues* (i.e. dimensions) that hinder SKOS vocabularies and supplies a framework for the automated assessment and correction of such issues. Those issues are grouped into three main quality categories, namely 'Labelling and Documentation issue', 'Structural issues' and 'Linked Data issues'. We want to point out that grouping issues in categories may facilitate the emergence of higher level views of the quality dimensions, allowing to highlight possible trade-offs between alternative objectives [33]. Leveraging on the IQ issues in [16] avoided us to specify 'another set' of IQ dimensions. In fact, we are more interested in discussing how the AHP may be useful to address the thesauri quality assessment than showing the appropriateness of new metrics. Due to its generality, the proposed approach may be applied even if the set of dimensions is different or if the assessment and correction automatic tools should change.

B. A Multi Criteria Decision Making approach to thesauri quality assessment

Due to the heterogeneity of the multiple IQ dimensions, the task of synthetizing an overall measure from the evaluation of the various dimensions is not, in principle, an easy one. Multi Criteria Decision Making techniques support such a "mixing apples with oranges" process. They regard the analysis of a set of various (finite or infinite) *alternatives*, namely the decision space, described in term of multiple *criteria*, aimed at deriving the ones better performing respect to the goal of the planning process [34]. To evaluate each alternative and to be able to compare it with others, the selection of criteria (aka attributes) is required to reflect the alternative performance in meeting the objective. Criteria represent the different dimensions from which the alternatives can be viewed [35]. Each criterion must be measurable to assess how well a particular option is expected to perform in relation to the criterion.

Multi Criteria Decision Making techniques may be applied to solve different classes of decision problems: choosing a single alternative; classifying alternatives into ordered predefined categories; ranking alternatives in a preference list [36]. According to this classification, the comparison of thesauri, the goal of the thesauri quality assessment, can exploit techniques supporting choice and/or ranking problems. The AHP [17] is particularly useful to supports decision-makers in structuring problem complexity and exercising judgment, allowing them to incorporate both objective and subjective considerations in the decision process.

The AHP methodology involves the execution of six phases: 1) selecting the criteria that characterize the decision problem alternatives and organizing them as a hierarchy; 2) pairwise comparing criteria according to user preference and achieving weights derivation; 3) evaluating or gathering the performance of each alternative with respect to each criterion; 4) scaling of criteria; 5) synthetizing and ranking the alternatives; 6) selecting the high ranking alternative(s).

Although the decision maker is the main actor involved, supporting tools[19] can simplify the human activity making the process more efficient. The availability of these tools allows to implement the thesauri quality assessment process in a semiautomatic way. We used the SuperDecisions[20] software that implements the AHP, as it granted us a free trial for one year. However, the same results are attainable using any other tool or even applying directly the methodology by means of any other mathematical software. The UML activity diagram in Fig. 1 summarizes the application of the six AHP phases to the documentation of the thesauri quality assessment process. The roles of the decision maker and of the supporting AHP tools are highlighted through corresponding diagram swimlanes. The rightmost swimlane accounts for the input and output quality documentation activities involved in publishing metadata. Though the six AHP phases are common to all possible application domains, in next section we summarize each phase by discussing its adoption to the thesauri quality assessment process. In particular, as a proof of concept, we focus on the thesauri quality assessment related to the maintenance of the linked thesauri of LusTRE.

IV. Overall Quality Assessment Of Linked Thesauri For The Environment

A. LusTRE framework of linked thesauri for the Environment

LusTRE is a linked thesaurus framework for the Environment. It has been developed within the EU funded project eENVplus concerning the deployment and integration of environmental services for advanced application within INSPIRE[21]. INSPIRE is the EU directive aiming to establish a Spatial Data Infrastructure (SDI) for Europe enabling the sharing of spatial data among public-sector organizations and facilitating their public access across Europe. It requires the deployment of geographic data in a standard way, and the provision of metadata with common nomenclature. LusTRE

[19] https://jyx.jyu.fi/dspace/handle/123456789/49477
[20] http://www.superdecisions.com/
[21] http://inspire.ec.europa.eu/

supports metadata provision by facing with cross-sectorial and cross-languages issues arising from data sharing.

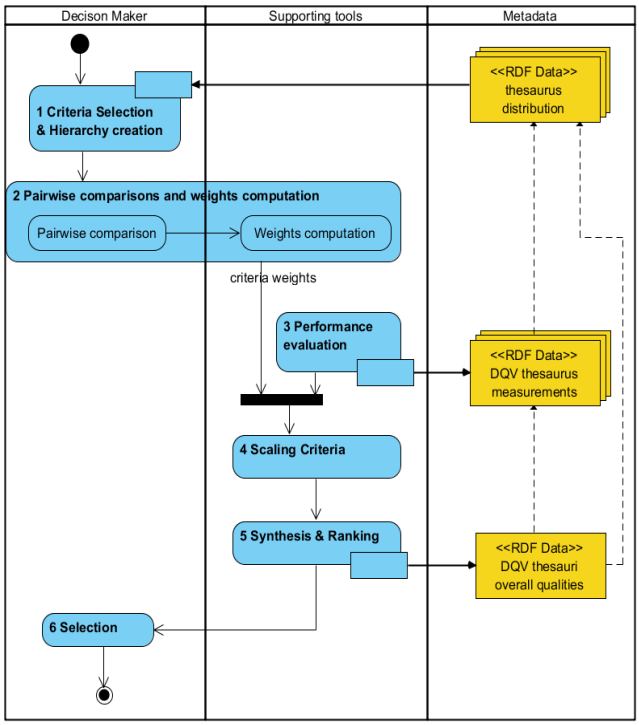

Fig. 1 Application of the six phases of AHP to the documentation of the thesauri quality assessment process.

Scientific observation and social data may be integrated with SDI for enhanced data analysis and scientific discovery. For instance, during environmental hazard (e.g. earthquake, floods, and landslides), there is an increasing social engagement from the Web and online social media, either commenting or communicating the environmental situation. The flourishing of monitoring networks integrating sensors and other data sources managed by volunteer's communities further amplifies such citizen involvement. The detected event could alert and trigger analysis processes on scientific observation data, or confirm results from scientific analysis [37]. The whole process combining Sensor Web, social data mining, and geoprocessing workflows for timely decision support may be accomplished by an SDI approach for Big Data analytics [38].

LusTRE provides wide shared standards and scientific terms for a common understanding of environmental data among the several communities operating in the Environmental sectors. It promotes the integration and jointly exploitation of different environmental KOSs terminologies. Based on the Linked Data principles, LusTRE provides a knowledge infrastructure of linked thesauri for the different cross-disciplines within the Environment. It can be used as one (virtual) integrated linked data source and a set of web services to exploit it within client applications for metadata editing and data search and indexing. The knowledge infrastructure consists of a set of SKOS terminologies related to different INSPIRE data themes. Their concepts are linked with the concepts of other vocabularies uploaded in LusTRE's SPARQL endpoint (e.g. GEMET, AGROVOC) or exposed in the Linked Open Data Cloud (e.g. EuroVoc, DBpedia). Currently, we are focusing on the framework maintenance, in particular on the quality issues affecting the publication and management of thesauri as Linked Data. This task involves the analysis and comparison of the thesauri quality to identify those requiring overriding technical improvement.

B. *Applying AHP to LusTRE*

Considering the UML diagram of Fig. 1, in the following, we briefly discuss the application of AHP to LusTRE thesauri.

1) Criteria selection and hierarchy creation. A set of criteria affecting LusTRE's thesauri (i.e. problem alternatives) is identified. The chosen criteria coincide with a subset of 14 of the 26 quality issues presented in [16], that we deemed as relevant for the maintenance activity, (listed in Table 1). Hierarchy organization helps the decision makers to organize a complex problem into its basic and simpler elements. This facilitates the assessment of the trade-off between criteria at the various level of the hierarchy at the basis of the AHP methodology. Accordingly, the selected criteria are hierarchy organized in a tree-like structure and grouped according to three categories: 'Labelling and Documentation issues', 'Structural issues' and 'Linked Data issues proposed in [16].

2) Pairwise comparison of criteria and weights computation. Reciprocal paired comparisons allow expressing judgments on the relative importance of each criterion by relating them to a scale of absolute numbers as defined by Saaty [17]. Operatively, the pairwise comparison made for each branch of each level of the hierarchy tree is mapped to square matrixes with the number of elements equal to the nodes at that branch. If an element K of the matrix is considered j times (with j an integer in the Saaty's scale) more important than an element Y, then it follows that Y is 1/j times as important as K. Based on the pairwise elicitation of relative importance of criteria given in matrix form, AHP allows estimating the criteria's weights (w_i). This can mainly be done using two methods: the logarithmic least squares method or the eigenvector method. The latter is advocated as more powerful as it allows dealing with inconsistencies that may arise from the elicitation process.

3) Performance evaluation. This phase requires the computation of performance values of each criterion according to every thesaurus (P_{ij}). As depicted in Fig. 1, in general, it is assumed that thesauri are published as RDF. We do not make assumptions about how the RDF is stored. We assume publishers have provided sufficient metadata to make the thesaurus accessible independently from how they have physically organized the distribution of their thesauri (e.g., distinct files, RDF stores, SPARQL endpoints). In the particular case of LusTRE's thesauri, the performance values are computed by applying the extended qSKOS directly to the RDF distribution of them. As results, the quality measurements for each criterion on each thesaurus are encoded in RDF according to DQV as shown in next Section. Each qSKOS outcome is represented as a DQV quality measure associated to

a dimension corresponding to the assessed criterion. This is made automatically by the modified qSKOS tool which outputs the results directly as DQV compliant documentation.

TABLE 1. THESAURI QUALITY ISSUES CATEGORY. ABSOLUTE ERRORS BY QSKOS AND SCALED SCORES (IN ROUND BRACKETS)

Quality issues	**EARTH**	**ThIST**	**GEMET**	**EuroVoc**	**AGROVOC**
N° Authoritative Concepts	14352	34155	5257	6883	32323
Labelling and Documentation issues					
Omitted invalid tags	0 (0)	0 (0)	3 (0,016)	240 (1)	55 (0,049)
Incomplete language coverage	461 (0,032)	24055 (0,705)	904 (0,172)	5173 (0,752)	32310 (1)
Inconsistent prefLabel	0 (0)	133 (1)	0 (0)	0 (0)	0 (0)
Disjoint label violation	69 (1)	1 (0,006)	3 (0,119)	0 (0)	6 (0,038)
Structural issues					
Cyclic hierarchical relations	0 (0)	9 (1)	0 (0)	0 (0)	3 (0,352)
Valueless associative relations	1124 (1)	2378 (0,889)	31 (0,075)	8 (0,015)	1671 (0,660)
Omitted top concept	1 (0,003)	1 (0,001)	1 (0,009)	139 (1)	0 (0)
Top concept having broader concept	0 (0)	0 (0)	0 (0)	0 (0)	4 (1)
Unidirectional related concept	0 (0)	39 (0,001)	0 (0)	15033 (1)	21351 (0,302)
Relational clashes	61 (1)	98 (0,675)	2 (0,089)	1 (0,034)	79 (0,575)
Mapping clashes	0 (0)	5 (0,296)	0 (0)	0 (0)	16 (1)
Linked Data issues					
Missing In-links	8530 (0,658)	30838 (1)	471 (0,099)	4439 (0,7143)	29111 (0,998)
Missing Out-links	8530 (0,659)	30821 (1)	472 (0,099)	4442 (0,715)	29111 (0,998)
Broken links	39 (0)	178 (0)	206 (0,002)	120.790 (1)	160 (0)
Overall Quality					
Synthesis scores	0,282	0.416	0,044	0,431	0,404
Ranks	4	2	5	1	3

4) Scaling criteria. The goal of scaling is to bring all criterion values into non-dimensional scores within the [0,1] interval, and thus make them comparable. Saaty's *Ideal* mode suggests to compare each performance value P_{ij} to a fixed benchmark, usually the maximum value achieved for criterion C_i amongst all the alternatives. The following normalization formula computes the score value S_{ij}:

$$S_{ij} = P_{ij}/max_j \text{ iff } max_j \neq 0; \quad S_{ij} = 0 \text{ iff } max_j = 0 \qquad (1)$$

with max_j the maximum value achieved for criterion C_j with respect to all thesauri. The higher the error, the greater the score achieved by a thesaurus for a given criterion, that will contribute to increase the ranking of the thesaurus when coupled with the computed weights. For each thesaurus, the scores obtained by (1) as listed in Table 1, are documented as DQV measurements with their dependencies of derivation.

5) Synthesis and ranking. The overall synthesis score for each thesaurus is obtained by adding the products of each criterion score S_{ij} with its associated weight w_i, across each branch of the hierarchy tree. This sum becomes the score value for the parent node directly above and the process is repeated at the next level of the hierarchy until the root node is reached. Higher synthesis scores mean more errors and higher thesauri ranks. Overall score is reported into DQV.

6) Selection. The higher-ranking thesauri are selected. In LusTRE thesauri, ranks reflect the amount of errors exposed. Therefore, higher ranking thesauri (i.e. EuroVoc, ThIST and AGROVOC) need to be fixed earlier than the lower ranking ones (see last row Table 1).

V. MAPPING THESAURI QUALITY ASSESSMENT TO DQV

We extended qSKOS tool, which finds quality issues in SKOS vocabularies, to encode its results in the DQV[22]. qSKOS quality issues can be considered as DQV quality dimensions, whilst the number of issues which occurs represent the metric deployed for each quality dimension. We have defined each qSKOS-related metrics and dimensions in proper namespace `qs`, which corresponds to http://w3id.org/quality/qskos/.

For example, the issue (i.e. dimension) and metric from the first row of Table 1 are mapped by defining `qs:omittedOrInvalidLanguageTags` as instance of `dqv:Dimension`, and `qs:numOmittedInvalidLangTags` as a new instance of `dqv:Metric`, as in the following RDF excerpt expressed in the Turtle Syntax[23]:

```
prefix dqv: <http://www.w3.org/ns/dqv#>
prefix qs: <http://w3id.org/quality/qskos/>
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
qs:omittedOrInvalidLanguageTags
   a dqv:Dimension;
   skos:prefLabel "Omitted Or Invalid Language Tags "@en ;
   skos:definition "Some controlled vocabularies contain literals in natural language, but without information what language has actually been used. Language tags might also not conform to language standards, such as RFC 3066."@en;
   dqv:inCategory qs:LabelingDocumentationIssues.
qs:numOmittedInvalidLangTags
   a dqv:Metric;
   skos:prefLabel "#Omitted/Invalid Language Tags"@en
   skos:definition "Number of omitted or invalid language tags"@en ;
```

[22] https://github.com/riccardoAlbertoni/qSKOS
[23] https://www.w3.org/TR/turtle/

```
   dqv:expectedDataType xsd:interger ;
   dqv:inDimension qs:omittedOrInvalidLanguageTags.
```

Once, dimensions and metrics are mapped, we can represent the actual values gauged by qSKOS instantiating the `dqv:QualityMeasurement` element.

Assuming the symbol ":" as a placeholder for an exemplificative namespace, and ":EARTh" and ":AGROVOC" as the URIs representing respectively the DCAT distribution for EARTh and AGROVOC, the following RDF triples represent their measures as in the first row of Table 1:

```
:exEARTh1 a dqv:QualityMeasurement ;
   dqv:computedOn :EARTh ;
   dc:date "2016-11-18"^^xsd:date ;
   dqv:value "0"^^xsd:integer ;
   dqv:isMeasurementOf qs:numOmittedInvalidLangTags.
```

and

```
:exAGROVOC1 a dqv:QualityMeasurement ;
   dqv:computedOn :AGROVOC ;
   dc:date "2016-11-18"^^xsd:date ;
   dqv:value "55"^^xsd:integer ;
   dqv:isMeasurementOf qs:numOmittedInvalidLangTags.
```

The `dc:date` Doubling Core element is deployed thus to document the quality assessment date.

Derived measures, i.e. the normalized scores, shown in Table 1 (in round brackets) under the qSKOS measures are encoded in analogy with the previous. For each of the previous measures we define the corresponded new scaled derived metric. For example, in the following, for `qs:numOmittedInvalidLangTags` we define `:scaledOmittedInvalidLangTagsForConcept` as a new derived metrics:

```
:scaledOmittedInvalidLangTagsForConcept
   a dqv:Metric;
   prov:wasDerivedFrom qs:numOmittedInvalidLangTags;
   skos:prefLabel "Scaled Omitted/Invalid Language Tags"@en ;
   skos:definition "Scaled number of omitted or invalid language tags for concept"@en ;
   dqv:expectedDataType xsd:decimal ;
   dqv:inDimension qs:omittedOrInvalidLanguageTags.
```

Accordingly, the associated scaled measurement for EARTh and AGROVOC are mapped as:

```
:exScaledEARTh1 a dqv:QualityMeasurement ;
   prov:wasDerivedFrom :exEARTh1 ;
   dqv:computedOn :EARTh ;
   dc:date "2016-11-18"^^xsd:date ;
   dqv:value "0"^^xsd:decimal ;
   dqv:isMeasurementOf
   :scaledOmittedInvalidLangTagsForConcept.
```

and

```
:exScaledAGROVOC1 a dqv:QualityMeasurement ;
   prov:wasDerivedFrom :exAGROVOC1;
   dqv:computedOn :AGROVOC ;
   dc:date "2016-11-18"^^xsd:date ;
   dqv:value "0.049"^^xsd:decimal ;
   dqv:isMeasurementOf
   :scaledOmittedInvalidLangTagsForConcept.
```

The overall quality measure for a thesaurus can be mapped into the DQV notation as a derived measurement computed on the derived qSKOS measurements.

We first map the metric that gauges the AHP outcomes when applied to the LusTRE thesauri with respect to a given context of use (e.g. the maintenance task) into the derived DQV `overallAHPQualityWithContextMaintenance` metric:

```
:overallAHPQualityWithContextMaintenance
   a dqv:Metric;
   prov:wasDerivedFrom
   :scaledOmittedInvalidLangTagsForConcept,…,
   :scaledBrokenLinks;
   skos:prefLabel "AHP Overall quality according to context LusTRE's Maintenance@en ;
   skos:definition "Describing the context LusTRE's Maintenance"@en ;
   dqv:expectedDataType xsd:decimal.
```

The associated overall quality measurements computed (by SuperDecisions) for EARTh and AGROVOC are mapped as:

```
:exoqEARTh a dqv:QualityMeasurement ;
   prov:wasDerivedFrom :exScaledEARTh1, …,
   …,exScaledEARTh14 ;
   dqv:computedOn :EARTh ;
   dc:date "2016-11-18"^^xsd:date ;
   dqv:value "0.282"^^xsd:decimal ;
   dqv:isMeasurementOf
   :overallAHPQualityWithContextMaintenance.
```

and

```
:exoqAGROVOC a dqv:QualityMeasurement ;
   prov:wasDerivedFrom :exnsAGROVOC1, …,
   …,:exnsAGROVOC14 ;
   dqv:computedOn :AGROVOC ;
   dc:date "2016-11-18"^^xsd:date ;
   dqv:value "0.404"^^xsd:decimal ;
   dqv:isMeasurementOf
   :overallAHPQualityWithContextMaintenance.
```

These RDF excerpts show a possible mapping of Table 1 into the DQV demonstrating how the quality assessment process can be integrated with an explicit metadata description.

VI. CONCLUSION

The paper focuses on SKOS-based Linked Data thesauri to easy cross-organization and cross-disciplinary management and use of Big Data/metadata. To effectively succeed the quality of the thesauri needs to be properly assessed and documented.

The paper combines a MADM-based methodology to assess the overall quality of linked thesauri, with the mapping

of the quality measures into a metadata vocabulary (DQV). The assessment exploits an overall quality measure based on the AHP, which takes into account both subjective and objective facets involved in the assessment process. This process is facilitated by the extension of the qSKOS tool that produces metadata documents, compliant with DCAT standard and W3C-DQV. The application of AHP to the linked thesauri deployed by LusTRE provides a proof of concept of the proposed approach demonstrating how quality can be documented in a specific testbed. Nevertheless, due to the generality of AHP and the DQV, such approach can be replicated in wider different contexts and be exploited for any kind of linked data provided that proper metrics are available. Future work analyzes how quality assessment and documentation can address data management efforts for ameliorating the utility of data [39] within different contexts.

REFERENCES

[1] D. Laney, "3D Data Management: Controlling Data Volume, Velocity and Variety", 6 Feb 2001, https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf .

[2] K. H. Leetaru, "A Big Data approach to the humanities, arts, and social sciences: Wikipedia's view of the world through supercomputing", Research Trends, 30 September 2012.

[3] S. K. Bansal and S. Kagemann, "Integrating Big Data: A Semantic Extract-Transform-Load Framework", Cover Feature Big Data Management, IEEE Computer Society, Issue No. 03 - Mar. (2015 vol. 48), pp 42-50, 2015.

[4] J. Hendler, "Broad data: Exploring the emerging web of data", in Big Data, vol 1, pp.18–20, February 2013.

[5] T. Berners-Lee, "Linked Data", http://www.w3.org/DesignIssues/LinkedData.html, 2009.

[6] I. Mitchell and M. Wilson, "Linked-data-connecting-and-exploiting-big-data," White paper Fujitsu, 2012.

[7] T. Heath and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space", Synthesis Lectures on the Semantic Web, Morgan & Claypool Publishers, 2011.

[8] C. Bizer, P.A. Boncz, M. L. Brodie, and O. Erling, "The meaningful use of big data: four perspectives - four challenges", SIGMOD Record, vol. 40, No. 4, pp.56-60, 2011.

[9] P. Hitzler and K. Janowicz, "Linked Data, Big Data, and the 4th Paradigm", in Semantic Web 4(3), pp.233-235, 2013.

[10] B. Farias Loscio, C. Burle, and N. Calegari, "W3C Data on the Web Best Practices", January 2017, W3C Recommendation.

[11] S. Neumaier, J. Umbrich, and A. Polleres, "Automated Quality Assessment of Metadata across Open Data Portals", in J. Data and Information Quality, Vol. 8 Issue 1, pp.2-29, 2016.

[12] A. Shiri, "Linked data meets big data: A knowledge organization systems perspective", 24th ASIS SIG/CR Classification Research Workshop, pp. 16–20, 2014.

[13] J. Chen and H. Yang, "From Data Reuse to Knowledge Reuse in Web Applications: A Survey", COMPSAC Workshops, pp. 174-179, 2016.

[14] A. Quarati, R. Albertoni, and M. De Martino, "Overall quality assessment of SKOS thesauri: An AHP-based approach", Journal of Information Science, 2016 (to be published).

[15] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked open data: a survey", Semant. Web J. 7(1), pp. 63–93, 2016.

[16] O. Suominen and C. Mader, "Assessing and improving the quality of SKOS vocabularies", Data Semant. J., vol. 3, pp. 47–73, 2013.

[17] TL. Saaty, The Analytic Hierarchy Process, McGraw-Hill, New York, 1980.

[18] R. Albertoni, M. De Martino, and A. Quarati, "Integrated Quality Assessment of Linked Thesauri for the Environment", EGOVIS 2016, pp. 221-235, 2016.

[19] A. Abecker, R. Albertoni, M. De Martino, P. Podestà, K. Schnitter, and R. Wössner, "Latest Developments of the Linked Thesaurus Framework for the Environment (LusTRE)", in ENVIROINFO 2015 Conference, Copenhagen Denmark , 2015.

[20] C. Mader, B. Haslhofer, and A. Isaac, "Finding quality issues in SKOS vocabularies," in Proceedings of the second international conference on theory and practice of digital libraries (TPDL 2012). LNCS, Springer, Heidelberg, vol. 7489, pp. 222–233, 2012.

[21] F. Maali and J. Erickson. "W3C. Data Catalog Vocabulary (DCAT)." January 2014. W3C Recommendation.

[22] R. Albertoni and A. Isaac. "Data on the web best practices: Data quality vocabulary", December 2016, W3C note.

[23] C. Batini and M. Scannapieco, Data and Information Quality. Dimensions, Principles and Techniques, Springer, Heidelberg, 2016.

[24] R. Albertoni and A. Gòmez-Pérez, "Assessing linkset quality for complementing third-party datasets", in Joint EDBT/ICDT 2013 Workshops, ACM, New York, pp. 52-59, 2013.

[25] D. Kontokostas, A. Zaveri, S. Auer, and J. Lehmann, "TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data", in 4th Conference on Knowledge Engineering and Semantic Web, 2013.

[26] J. Debattista, A. Sören, and C. Lange, "Luzzu - A Methodology and Framework for Linked Data Quality Assessment", Data and Information Quality J., Vol. 8(1), pp.1-32, 2016.

[27] D. Kless and S. Milton, "Towards quality measures for evaluating thesauri", in 4th metadata and semantics research conference, Comm. in computer and information science 108, pp. 312–319, 2010.

[28] R. Albertoni, M. De Martino, and P. Podestà, "A Linkset Quality Metric Measuring Multilingual Gain in SKOS Thesauri", in 2nd Workshop on Linked Data Quality, colocated in ESWC 2015, 2015.

[29] F. Radulovic, N. Mihindukulasooriya, R.García-Castro, and A. Gómez-Pérez, "A comprehensive quality model for Linked Data", to apper in Semantic Web Journal, http://www.semantic-web-journal.net/content/comprehensive-quality-model-linked-data-1

[30] J. Lehmann, R. Isele, M. Jakob, et al. "DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia", Semantic Web Journal, vol. 6 No. 2, pp 167–195, 2015.

[31] M. Brümmer, C. Baron, I. Ermilov, M. Freudenberg, D. Kontokostas, and S. Hellmann: "DataID: towards semantically rich metadata for complex datasets", SEMANTICS 2014, ACM, pp.84-91, 2014.

[32] A. Even and G. Shankaranarayanan, "Value-Driven Data Quality Assessment", in Proceedings of the 10th International Conference on Information Quality, Cambridge, 2005.

[33] D.P. Ballou and H.L. Pazer, "Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts", IEEE Trans. Knowl. Data Eng. 15(1): pp.240-243, 2003.

[34] E. Kazimieras Zavadskasa, Z. Turskisa, and S.Kildienė, "State of art surveys of overviews on MCDM-MADM methods" Technological and Economic Development of Economy, 20(1), pp.165-179, 2014.

[35] E.Triantaphyllou, B.Shu, S.N. Sanchez, and T. Ray, "Multi-Criteria Decision Making: An Operations Research Approach," Encyclopedia of electrical and electronics engineering 15, 175-186,1998.

[36] B. Roy, "The optimisation problem formulation: criticism and overstepping" The Journal of the Operational Research Society 32(6), pp. 427-436, 1981.

[37] T. Bedrina, A. Parodi, A. Quarati, and A. Clematis, "ICT approaches to integrating institutional and non-institutional data services for better understanding of hydro-meteorological," Nat. Hazards Earth Syst. Sci., 12, pp. 1961–1968, 2012.

[38] P.Yue, C. Zhang, M. Zhang, X.Zhai, and L.Jiang, "An SDI approach for big data analytics: The case on sensor web event detection and geoprocessing workflow," IEEE J. Sel. Top. Appl. Earth Observ., vol.8, pp. 4720–4728, 2015.

[39] A. Even and G. Shankaranarayanan, "Understanding impartial versus utility-driven quality assessment in large datasets", In: ICIQ, pp 265–279, 2007.