Article

JIS

Journal of Information Science
1–20
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

Overall quality assessment of SKOS thesauri: an AHP-based approach

Alfonso Quarati
Institute for Applied Mathematics and Information Technologies, National Research Council, Italy

Riccardo Albertoni
Institute for Applied Mathematics and Information Technologies, National Research Council, Italy

Monica De Martino
Institute for Applied Mathematics and Information Technologies, National Research Council, Italy

Abstract

The paper proposes a methodology for a Thesauri Quality Assessment that supports decision makers in selecting thesauri by exploiting an overall quality measure. This measure takes into account the subjective perceptions of the decision maker according to the reuse of thesauri in a specific application context. The Analytic Hierarchy Process methodology is adopted to capture both subjective and objective facets involved in the Thesauri Quality Assessment, thus providing a ranking of the thesauri assessed. Our methodology is applied to a set of thesauri by using user-driven application contexts. A step-by-step explanation of how the approach supports the decision process in the creation, maintenance and exploitation of a framework of linked thesauri is provided.

Keywords

SKOS; thesaurus; data quality; overall quality measure; Analytic Hierarchy Process

1. Introduction

Terminological resources such as thesauri, or controlled vocabularies [1], are instances of Knowledge Organization Systems (KOSs) used within information retrieval systems to provide a uniform description for information resources as well as in search and discovery processes [2].

Thesauri enable users to share and agree on scientific/technical terms and to express them in multiple languages. Several thesauri have been deployed by domain-specific communities embodying various points of view based on alternative conceptualizations. Their development depends on specific goals and involves different levels of abstraction and details. Many organizations and projects [3-5] are now interested in publishing their terminological resources on the web according to the Linked Data (LD) initiative, namely the Simple Knowledge Organization System (SKOS) [6].

SKOS was designed to provide a common data model for KOS to be published according to LD principles [7]. LD exploits the web to connect data and is aimed at transforming the web into a global knowledge base [8]. LD technology makes thesauri available not just as isolated islands but as part of a federation of interconnected thesauri, referred to as Thesaurus Framework (TF), providing extensive sets of terminologies for different communities working on the same domain [9, 10].

The identification and selection of reusable thesauri is critical. Recent works have addressed the reusability of vocabularies in terms of domain coverage, multilingual support, and technological aspects [11, 12]. However, the quality issue is critical for thesauri reuse, and is strongly affected by the constraints and particular features of the application context.

The problem of the quality assessment of LD resources has focused above on defining metrics for measuring the quality of datasets according to specific dimensions [13]. Suominen and Mader [14] presented an extended systemization of

Corresponding author:
Alfonso Quarati, Institute for Applied Mathematics and Information Technologies, National Research Council, Italy
Email: quarati@ge.imati.cnr.it

common quality issues in SKOS vocabularies and introduced automated assessment and correction tools aimed at fixing a subset of the quality issues identified. These issues are considered from an objective point of view, disregarding any subjective evaluations from specific real case scenarios. They thus focus on highlighting syntactic (i.e. with respect to the SKOS and LD specifications) or lexical issues (e.g. extra white space in labels). Most of these works generally focus on sets of single quality measures, whereas the selection of reusable KOS may benefit from the attribution of an overall measure that accounts for the quality of thesaurus as a whole. This overall measure is not generally univocal, but depends on the specific application context.

This paper proposes a Multiple Criteria Decision Making (MCDM) based methodology for assigning an aggregated quality value to a thesaurus according to the requirements of the specific application context. We deal with three application contexts regarding the management of a linked thesauri framework. We introduce the Thesauri Quality Assessment (TQA) aimed at selecting SKOS thesauri through the exploitation of an overall quality measure. We use the Analytic Hierarchy Process (AHP) proposed by Saaty [15] in order to capture both subjective and objective facets involved in the TQA, and so that we can rank the thesauri assessed. We provide a step-by-step explanation of how this approach supports the decision tasks involved in the creation, maintenance and exploitation of a TF. As a reference scenario we consider the thesauri and quality measures presented by Suominen and Mader [14], along with a further quality indicator regarding the number of thesaurus authoritative concepts.

The rest of the paper is organized as follows. Section 2 overviews the background and related work. Section 3 introduces the TQA process, as well as the reference scenario and the application contexts. Section 4 describes the MCDM-based methodology in general, and AHP in particular. Section 5 illustrates the application of the methodology to rank the thesauri of the reference scenario according to different application contexts and analyses the results. Section 6 discusses the impact of the proposed approach, and Section 7 presents the conclusions and future work.

2. Background and related work

Information Quality (IQ) is a very long-established issue, where questions of authority, veracity and reliability have been raised since the establishment of printing in Europe. Technological developments have given further significance to IQ [16]. From database systems to multi-sourced and multimedia information delivered by Web 2.0 technologies, poor data quality affects information integration, data sharing, information retrieval, information extraction, and knowledge discovery [17]. In the following sections we discuss various IQ issues and highlight recent works regarding SKOS thesauri quality.

2.1. You cannot control what you cannot measure

Several works [18-20] refer to IQ as “fitness for use”, adopting Juran's broad definition [21]. IQ is related to those features of products or services that meet customer needs and thereby provide customer satisfaction. As noted by Tayi and Ballou [19], this definition expresses a relative and task-dependent concept: a satisfactory quality level for one user may be insufficient or inappropriate for another.

According to Bizer and Cyganiak [22] IQ has two facets: it is task-dependent and it is subjective. This definition correctly highlights the importance of recognizing the scope of an IQ process. However, it also needs to specify how IQ can be obtained, given the context of use and the relative user perspective. Naumann [23] argues against “fitness for use” or definitions of IQ such as “meets information consumers’ needs” [24] and “user satisfaction” [25]. According to Naumann, these definitions are just as non-operational as Pirsing's: “Even though quality cannot be defined, you know what it is” [26]. He therefore envisages “quality as an aggregated value of multiple IQ-criteria” and maintains that this definition:

> is flexible regarding the application domain, the sources, and the users, because the selection of criteria can be adapted accordingly

concluding that:

> assessing scores for certain aspects of information quality and aggregating these scores is easier than immediately finding a single global IQ-score

Similarly, Tayi and Ballou [19] claim that data quality problems can only be addressed by a proper understanding of their multiple attributes or dimensions. Various attempts to define data quality dimensions have been proposed including availability, accuracy, timeliness, completeness, relevancy, consistency [27,28]. Wang et al. [18] solicited information from users regarding various data quality descriptors, and derived 20 categories which were further classified into four data quality groups: intrinsic, contextual, representational, and accessibility. In another work, Naumann [29] quoted DeMarco's famous assertion [30]:

> You cannot control what you cannot measure.

and re-affirmed the importance of defining measurable quality dimensions, according to the problem at hand. He also stressed the importance of clearly evaluating the aim of an IQ process and questioned why IQ should be assessed. In response he provided several possible answers:

> estimating quality relevance and significance; need for improvement; and in case of improvement: cost-benefit ratio

An IQ problem can thus be correctly assessed if both subjective and objective perspectives are taken into account, and provided that it is actually possible to measure them all.

2.2. SKOS quality

Specific aspects of linked dataset quality have been addressed by proposing sets of evaluation metrics and methodologies. An IQ metric is used to measure an IQ dimension. In general, more than one metric may be used to assess one dimension. As a quality dimension can often be rather abstract, metrics rely on IQ indicators. Indicators are characteristics of datasets (e.g., pieces of dataset content, dataset meta-information, human ratings) regarding the suitability of a dataset for an intended use. The IQ assessment is used to evaluate whether a piece of information meets a specific consumer's information needs [22].

Guéret et al. [31] examined five network measures, as a way of capturing the quality of interlinks amongst datasets. The authors affirmed that such measures are only partially effective at detecting bad links and concluded that more tailored network measures are needed. They also recognized the need to analyse the interplay of different measures and the interpretation of the combined results. Albertoni and Gómez-Pérez [32] presented a specific quality dimension for interlinking, aimed at assessing the linkset quality for complementing third party datasets. Completeness is defined as the degree to which there are no missing links in the linkset.

Other proposals rely on the adoption of crowdsourcing to assess the correctness of datasets [33, 34]. Tools are proposed for delivering micro-tasks (i.e. sets of RDF triples possibly augmented by "surrounding" information) to heterogeneous crowds. After the crowd have carried out their tasks, the overall dataset correctness is assessed. Alternatively, RDF triples deemed as incorrect by the majority of the crowd are removed. Depending on the size of the dataset, this process may be extremely time-consuming, and costly if the crowd participants are also paid.

Bizer and Cyganiak [22] presented a Web Information Quality Assessment (WIQA) framework for web information consumers. WIQA applies different filtering policies relying on complex metadata such as provenance chains and background information on providers. It proposes a policy language and deploys an interpretation engine to explain why the information satisfies a specific policy. However, WIQA requires a considerable amount of user involvement. Debattista et al. [35] presented a quality assessment framework for linked open data (LUZZU), which provides quality metadata and quality problem reports for data cleaning. LUZZU is extensible, enabling third party metrics to be plugged into the framework. The authors introduce 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 a dataset is ranked by using a weighted sum on all the metrics. However, no major insights are given regarding the underlying methodology supporting this ranking, whereas in this paper we explain our rationale in detail.

Kless and Milton [36] suggested a range of abstract measurement constructs based on quality notions in the thesaurus literature and inferences from other related literature. The purpose of the measurement constructs (i.e. dimensions) is to support the evaluation of thesauri, however as reported by the authors, such constructs are solely based on a theoretical analysis. They highlight the need to operationalize the measures and then refine them by applying them to real cases. Moreover, their constructs do not address the quality dimensions related to multilingual thesauri.

Albertoni et al. [37] extended the quality dimension for interlinking presented by Albertoni and Gómez-Pérez [32]. They describe a linkset metric to estimate the ability of a linkset to enrich a dataset with new property values. Sabbah et al. [38] analysed the effect of the thesaurus size on the quality of schema matching. However, they did not examine the measurement and assessment of the thesaurus quality.

To date the most complete IQ work related to SKOS thesauri is provided by Suominen and Mader [14], who introduce a set of 26 quality issues, defined as computable functions exposing potential quality problems. Using the open source qSKOS tool [39], they analysed a corpus of 24 vocabularies, assessing their quality with respect to the 26 issues. They used Skosify [40] to automatically correct a subset of these issues. They present several facets of the vocabulary quality assessment problem and provide useful observations, recommendations and best practices. Despite not aiming to address an overall measure of the quality of a vocabulary, citing the work of Pipino et. al. [41], they maintain that:

> data quality should involve both "subjective perceptions of the individuals" and "objective measurements based on the data

and state that their work focuses on the second aspect. By contrast, our proposal aims to conciliate both these polarities by adopting an approach (i.e. AHP) that also considers the subjective side of the assessment process.

Using an MCDM methodology to assess knowledge resources for reuse in an ontology is discussed by Jiménez et al. in [42]. Lozano-Tello and Gómez-Pérez applied AHP to measure the suitability of existing ontologies, regarding specific user requirements [43]. Other works such as Guarino and Welty [44] and Hwang et al. [45] discuss ontology engineering methods which can be exploited to address semantic and conceptual consistency. In contrast our paper presents a methodology for the aggregation of quality measures according to different contexts tailored to user needs.

3. Thesauri Quality Assessment

A crucial problem for thesaurus developers is how to evaluate and select the most feasible thesauri in a corpus for reuse in a new application context. Feasibility is strictly related to the concept of quality [14]. Considering SKOS thesauri, several issues may hinder the quality of a thesaurus and therefore its reuse. In Section 2, we examined proposals aimed at separately evaluating the quality of various dimensions of datasets. Apart from Debattista et al. [35], these works do not address the problem of the actual quality of datasets as a whole. This paper aims to assess the overall quality of SKOS thesauri according to an application context and user needs.

In this paper the term Thesauri Quality Assessment refers to the selection of thesauri based on an overall quality measure. Our aim is to present a methodology that supports the TQA based on a well-founded decision making technique, which provides a rank for each thesaurus under evaluation. This rank is obtained by appropriately aggregating several IQ dimensions. A set of IQ dimensions is often assessed from an "objective" perspective, without considering the "subjective" point of view of the expert. Subjectivity entails prioritizing the importance of one dimension with respect to another, and providing a judgment on dimensions that cannot quantitatively be measured (by a procedure) and that require a qualitative assertion regarding their importance in terms of a given context. The approach we adopted is aimed at conciliating these two aspects. The TQA thus adheres greatly to the specific application, while at the same time being objectively grounded on computable metrics.

In the following sections we describe the reference scenario and the application contexts as a framework for our methodology.

3.1. Reference scenario

Our proposal is based on the work of Suominen and Mader [14] who provide a thorough discussion of the quality issues that impact on SKOS vocabularies and a framework for the automated assessment and solution to such issues. Their work gathers a representative data set of SKOS vocabularies, in order to ensure a wide coverage of domains. It considers vocabularies in each of the seven categories of the linked open data cloud domain classification (http://wifo5-03.informatik.uni-mannheim.de/lodcloud/state/#domains). For each domain, one small (up to 3,000 concepts), one medium-size (3,001–10,000 concepts) and one large (more than 10,000 concepts) SKOS vocabulary, is selected. A set of 26 issues affecting SKOS vocabulary quality, are considered and checked with qSKOS and Skosify. Many of these issues are grounded on, or at least refer to, thesaurus standards (e.g. ISO 25964). Issues are grouped into three categories, namely 'Labelling and Documentation', 'Structural' and 'Linked Data'.

We aim to provide an aggregate measure of the overall thesaurus quality by applying the AHP in the TQA process. As our goal was not to define new IQ indicators or metrics, we leveraged on a subset of the issues presented in Suominen and Mader [14]. Our reference scenario considers 20 of the 26 issues. It omits four issues not checked (or only partly) by the qSKOS tool, and another two with extremely poor values for all thesauri (i.e. approaching zero errors). Notice that with respect to the IQ terminology reported in Section 2, the term 'issue' (i.e. computable functions exposing potential quality problems) seems to incorporate both the dimensions (even more precisely the indicator) and the metric concept. We thus use these terms interchangeably.

In addition to the three issue categories, we consider a fourth category namely '1 Authoritative Concepts'. This is characterised by a unique element, '1.1 Authoritative Concepts', corresponding to the number of concepts identified by URIs in the thesaurus namespace. This is not an issue itself; rather it may be regarded as a quality indicator, according to the following rule of thumb: 'Not too few / Not too many' concepts should be present in a 'good' thesaurus. In fact, if a thesaurus has few concepts it will not provide the terminological coverage required to properly support the indexing and browsing of a large number of resources. On the other hand, thesauri exhibiting a huge amount of concepts are usually generated by collaborative editing or by automatic techniques, which may affect the overall terminological precision and coherence. Table 1 highlights the issue categories of the reference scenario - further details of each issue are provided by Suominen and Mader [14].

Grouping issues into categories reveals higher-level views of the quality issues, thus highlighting possible trade-offs between alternative objectives. We numbered the four categories and the 21 quality indicators accordingly by prefixing each issue with the number of category it belongs to. Although not mandatory, this helps visually in applying the AHP to the problem at hand, especially in the organization of issues hierarchically.

Note that the focus of this paper is not to identify the best set of quality measures. In fact, the proposed approach may be applied even if the set of indicators is different or if the assessment and automatic correction tools change.

Suominen and Mader [14] explicitly asserted that they:

> did not assign grades of severity to the issues, because such a judgement is highly dependent on the context and intended application of the vocabulary

Our proposal aims to show how, through a careful analysis of the TQA process, and through a suitable adoption of the AHP, it is possible to assign issues with different priorities according to the specific context.

Table 1. Description of quality issue categories of the reference scenario

Category	Quality Issues
1 Authoritative Concepts	Authoritative Concept is a quality indicator that estimates a thesaurus according to the rule of thumb: 'Not too few / Not too many. Thesauri are mapped into the following ordered set of categorical values: Unsatisfactory, Sufficient, Average, Good, Very Good.
2 Labelling and Documentation	Labelling and documentation criteria check that language tags are correctly indicated in thesaurus labels (2.1 Omitted or invalid language tags), that labels and definitions are provided for every concept and in the same set of languages (2.3 Undocumented concepts and 2.2 Incomplete language coverage respectively), that SKOS consistency axioms are not contradicted (2.4 Overlapping labels, 2.5 Inconsistent preferred labels, 2.6 Disjoint label violations) and that extra spaces are not erroneously included in labels (2.7 Extra whitespace in labels).
3 Structural	Structural criteria check for single isolated concepts or groups of isolated concepts (3.1 Orphan concepts and 3.2 Disconnected concept clusters respectively), for cycles in hierarchical relations (3.3 Cyclic hierarchical relations), for associative relations between concepts with a common broader concept (3.4 Valueless associative relations), for transitive relations in SKOS, such as broaderTransitive and narrowerTransitive, which are meant to be inferred by the vocabulary consumer and should not be directly used (3.5 Solely transitively related concepts), for top concepts that are not related via the hasTopConcept property (3.6 Omitted top concepts), for top concepts with broader concepts (3.7 Top concepts having broader concepts), for asymmetric associative properties (3.8 Unidirectionally related concepts), for disjointedness among the transitive closures of associative and hierarchical relations (3.9 Relation clashes) and for the disjointedness among mapping relations (3.10 Mapping clashes).
4 Linked Data	Linked data criteria identify authoritative concepts that have no links from or to other resources on the Web (4.1 Missing In-links and 4.2 Missing Out-links) and links between resources that return HTTP error responses or no response at all when being dereferenced (4.3 Broken links).

3.2. Application contexts

We introduce three application contexts used as a proof of concept to demonstrate the MCDM methodology as a means for capturing both the subjective and objective facets of the TQA. They are derived from our past experience acquired within the EU project eENVplus (http://www.eenvplus.eu). In eENVplus we were faced with the problem of reusing different thesauri for cross-sectorial data sharing within a European Spatial Data Infrastructure. We developed a framework of multilingual linked thesauri, namely LusTRE, to provide a wider terminology and a set of services to be exploited by metadata editing and data discovery (e.g. browsing and indexing) at the European level [9,10]. Currently it is exploited through a set of web services by EUOSME metadata editor (http://showcase.eenvplus.eu/client/editor.htm), and QSphere plugin (https://www.fgdc.gov/iso-metadata-editors-registry/editors/11).

A framework of interlinked thesauri provides extensive sets of terminologies for different communities working on the same domain [46, 47, 48]. The main tasks related to the life-cycle of such a framework are: (i) the selection of existing thesauri, with an emphasis on multilingualism, openness and usability features; (ii) the publication of thesauri as LD; (iii) the creation of interlinking between the published thesauri, as well as those already made available as LD by third-parties; (iv) the maintenance of the framework after its publication; (v) the framework exploitation, for example, in order to support the metadata compilation and the browsing and indexing of resources. These tasks involve thesaurus developers and experts from the domain as well as professionals in the LD sector (i.e. TF manager).

From our experience in the implementation and exploitation of a TF, we understood the importance of assessing the thesauri quality to determine their suitability according to distinct application contexts. In this paper, we focus on the following application contexts, which are further detailed in Section 5.1:

- AC1, thesauri selection for the creation of a TF, aimed at identifying the most suitable thesauri to be included in a framework. This is done by considering the relevant quality issues from the point of view of the thesaurus developers and domain-specific experts.
- AC2, TF maintenance, aimed at periodically evaluating the quality of the linked thesauri after their inclusion in the framework. The TQA compiles a list of thesauri majorly affected by errors, enabling a TF manager to fix them.
- AC3, thesaurus based indexing, which assesses thesauri to be exploited for indexing/query expansion. The TQA checks those quality issues that are more likely to affect the indexing/query expansion process.

In Section 5 we supply a step-by-step description of how to apply AHP to these application contexts.

4. MCDM-based methodology

In this section we provide an overview of the MCDM methodology along with a step-by-step introduction to AHP.

4.1. MCDM overview

The TQA relies on the aggregation of multiple quality dimensions thus supplying thesauri consumers with an overall quality measure, on the basis of which they can select the thesauri that best fit their needs. However, due to the heterogeneity amongst multiple IQ dimensions, it is not easy to obtain an overall measure that synthesizes the evaluation of the various dimensions.

To support the “mixing apples with oranges” process, typical of this decision task, we consider a well consolidated approach named Multi Criteria Decision Making (MCDM). According to the MCDM International Society [49]:

> MCDM can be defined as the study of methods and procedures by which concerns about multiple conflicting criteria can be formally incorporated into the management planning process

MCDM methods regard the analysis of a set of various (finite or infinite) alternatives, namely the decision space, described in terms of multiple criteria, aimed at deriving those that best perform with respect to the goal of the planning process. In the last few decades, thousands of papers have been published on MCDM methods, their development and application in different fields [50]. These methods propose procedures that, starting from the information available describing a given problem, evaluate the various alternatives (options) in order to fulfil a target objective. In order to evaluate each alternative and to compare it with others, the selection of criteria (attributes) is required to reflect alternative performances in meeting the objective. Different criteria represent the dimensions from which the alternatives can be viewed [51]. Each criterion must be measurable in order to assess how well a particular option is expected to perform in relation to the criterion. Criteria can be measured in cardinal numbers (e.g. price, number of drawbacks), in binary terms (e.g. a tick indicates the presence of a particular feature), or in qualitative terms (e.g. ‘good’, ‘insufficient’).

The most widely adopted MCDMs are: TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) [52], SAW (Simple Additive Weighting), AHP [15], and the family of methods originated by ELECTRE (ELimination and

Choice Expressing REality) [53]. In [54] Roy classifies a multi-attribute decision problem as: a) Choice problem: to select a single best alternative or to reduce the group of alternatives to a subset of equivalent or incomparable alternatives; b) Sorting problem: alternatives are sorted into ordered predefined categories; c) Ranking problem: alternatives are ordered in a decreasing preference list; d) Description problem: the goal is to describe alternatives and their consequences. This classification helps decision makers to identify the most suitable MCDM technique to address the specific formulation of a given problem.

TQA can thus leverage on techniques supporting choice and/or ranking problems. As observed by Ishizaka et. al. in [55], AHP is particularly useful for these kinds of problems. Its ability to structure a problem into a hierarchy of different levels (goal, criteria, sub-criteria and alternatives) fits well with our description of the TQA characterized by grouping the quality issues into four categories. We also adopted AHP due to its high rate of success in various fields [56].

4.2. The Analytic Hierarchy Process

AHP supports decision-makers in structuring problem complexity and exercising judgment, thus incorporating both objective and subjective considerations into the decision process [57]. AHP is made up of six phases which are common to all applications: 1) identifying the criteria that characterize the alternatives of the decision problem and organizing them into a hierarchy; 2) comparing pairwise criteria according to user preferences and obtaining criteria weights; 3) evaluating or gathering the performance of each alternative with respect to each criterion; 4) scaling of criteria; 5) synthesizing and ranking alternatives; and 6) selecting the best alternative(s). In Section 5 we show how AHP is adopted in the TQA process by considering our three application contexts.

The overall TQA process is shown in the UML diagram in Figure 1, in which the role of the decision maker (i.e. the thesaurus developer or the TF manager) and the supporting AHP tool are highlighted through corresponding diagram swimlanes. Although AHP can be applied without any software, MCDM tools [58] simplify the process greatly, saving time and making it more efficient. AHP tools enable the TQA to be implemented in a semi-automated way. The decision maker chooses the criteria and their hierarchical organization, as well as their pairwise comparison which is also usually supported by the AHP software. Phase 2 consists of two steps: the decision maker compares criteria and the tool computes the weights (the second activity box in Figure 1 crosses the border of the two swimlanes). The rest of the work is automatically executed by the software which returns the ranked list of m alternatives (i.e. thesauri), from which the decision maker can select the one(s) fitting his/her needs.

4.2.1. Phase 1: Criteria selection and hierarchy creation

This phase generally starts with the selection of the set of n criteria C_j (j=1,…,n) which, depending on the decision maker's requirements, qualify the specific decision problem, and characterize each of the m alternatives $\mathcal{T}_i$ (i=1,…,m). A generic TQA process bases this selection on the 21 issues (see Section 3.1). The hierarchical organization helps the decision makers to organize a complex problem into its basic and simpler elements [59]. This decomposition sets the goal (i.e. TQA) of the analysis at the top of the hierarchy, i.e. in our case the selection of thesauri with the highest quality. Depending on the complexity of the problem, as understood by the decision maker, the criteria can be further refined into sub-criteria, and so on.

4.2.2. Phase 2: Pairwise comparison of criteria and weights computation

Reciprocal paired comparisons express judgments on the relative importance of each criterion and sub-criterion and automatically link them to a numerical scale of absolute numbers (in the integer interval 1-9) defined by Saaty [60]. This phase establishes the 'local' priorities of the elements in each branch of the tree with respect to their parent node. The hierarchical structure is created for each branch of each level, using square matrixes with an equal number of elements to the nodes at that branch. If an element X of the matrix is considered j times (with j as an integer in the Saaty scale) more important than an element Y, then it follows that Y is 1/j times as important as X. Based on the pairwise elicitation of the relative importance of criteria given in matrix form, AHP computes the criteria's weights w_j. This is carried out using the logarithmic least squares method or the eigenvector method. According to Saaty and Vargas the eigenvector method is more powerful when dealing with inconsistencies that arise from the pairwise elicitation [61]. Criteria priorities can generally be provided by an individual or a group of people who discuss and achieve a consensus.

Phases 1 and 2 are fundamental to the adoption of AHP. They express the knowledge, interest and expectations of the human experts involved in the TQA process. They depend on the active role of the decision makers and on the application context. Both phases are facilitated by tools that enable decision makers to provide criteria information for the subsequent phases.

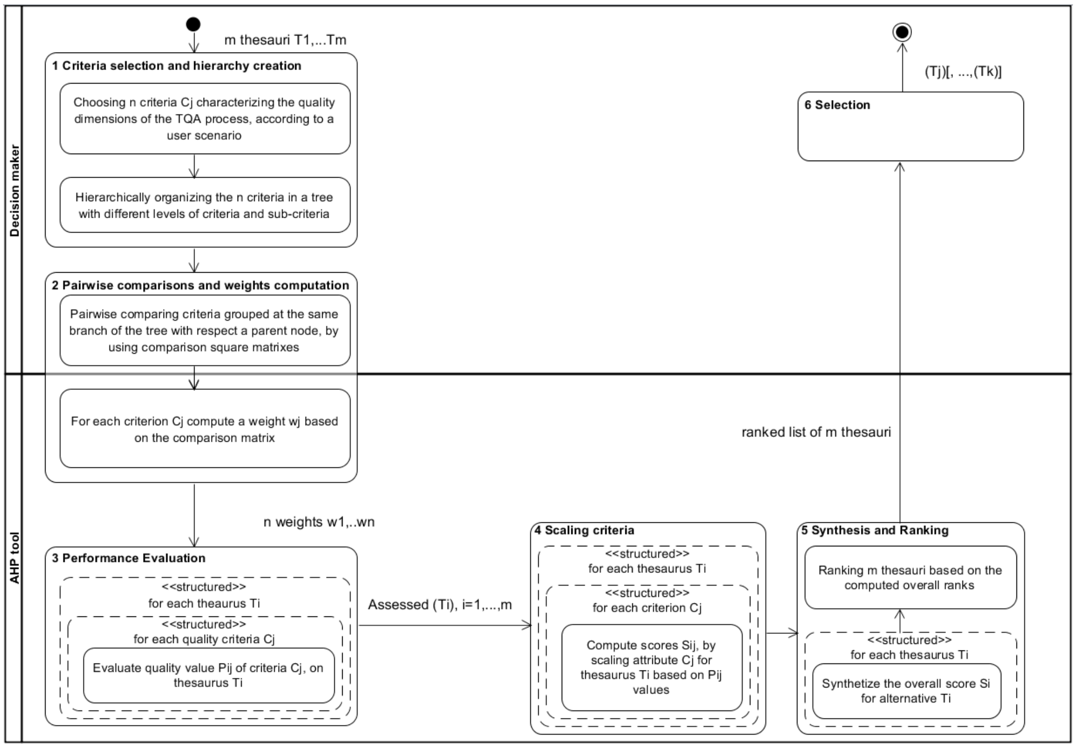

Decision maker
AHP tool
m thesauri T1,...Tm
1 Criteria selection and hierarchy creation
Choosing n criteria Cj characterizing the quality dimensions of the TQA process, according to a user scenario
Hierarchically organizing the n criteria in a tree with different levels of criteria and sub-criteria
2 Pairwise comparisons and weights computation
Pairwise comparing criteria grouped at the same branch of the tree with respect a parent node, by using comparison square matrixes
For each criterion Cj compute a weight wj based on the comparison matrix
n weights w1,..wn
3 Performance Evaluation
<<structured>>
for each theaurus Ti
<<structured>>
for each quality criteria Cj
Evaluate quality value Pij of criteria Cj, on thesaurus Ti
Assessed (Ti), i=1,...,m
4 Scaling criteria
<<structured>>
for each thesaurus Ti
<<structured>>
for each criterion Cj
Compute scores Sij, by scaling attribute Cj for thesaurus Ti based on Pij values
5 Synthesis and Ranking
Ranking m thesauri based on the computed overall ranks
<<structured>>
for each thesaurus Ti
Synthetize the overall score Si for alternative Ti
ranked list of m thesauri
6 Selection
(Tj)[, ...,(Tk)]

Figure 1. Thesauri Quality Assessment: the role of the decision maker and the support of an AHP tool.

4.2.3. Phase 3: Performance evaluation

In a generic AHP process, the overall performance of each alternative is a function of the performance of *n* criteria, whose values may either be derived by exogenous factors (e.g. natural phenomenon, the cost or monetary value of a good or service), or obtained by some procedural algorithm or formula, or established by the subjective perspective (either qualitative or quantitative) of the decision maker. Considering the TQA process, we assumed that the quality performance $\mathcal{P}_{ij}$ value for criterion $\mathcal{C}_j$, with respect a given alternative $\mathcal{T}_i$, (except for the '1.1 Authoritative Concepts' criterion) can be computed using a tool such as qSKOS.

4.2.4. Phase 4: Scaling criteria

Before comparing the *m* alternatives, the original performance values of each criterion measured according to different scales need to be transformed into relative values called scores. As performance values can be provided in numeric, descriptive or categorical data, the scaling of criteria solves the problem of different ranges and units of performance values. Naumann [23] noted that the goal of scaling is to bring all criterion values into non-dimensional scores within the [0,1] interval, thus making them comparable. To achieve this normalization task, Saaty and Vargas proposed several methods [62]. The so-called Ideal mode compares each performance value $\mathcal{P}_{ij}$ to a fixed benchmark, for example the performance value of the best alternative under criterion $\mathcal{C}_j$. In its general application, the Ideal mode computes the score values $\mathcal{S}_{ij}$ by dividing $\mathcal{P}_{ij}$ by the maximum value achieved for criterion $\mathcal{C}_j$ amongst all the alternatives.

4.2.5. Phase 5: Synthesis and ranking

Once scores have been computed for each criterion of each alternative, they are combined with the criteria weights derived from pairwise comparisons (Phase 2) to determine the overall synthesis score $\mathcal{S}_i$ for each alternative $\mathcal{T}_i$. $\mathcal{S}_i$ is computed by adding the products of each criterion score $\mathcal{S}_{ij}$ with its associated weight w_j, across each branch of the hierarchy. This sum is the score value for the parent node and the process is repeated at the next level of the hierarchy until the root node is reached. Based on the synthetized scores, an overall ranking is obtained by sorting the alternatives. These ranks account for the level of fulfilment of each alternative in achieving the goal.

4.2.6. Phase 6: Selection

The final AHP phase is to choose the best alternative or select from the highest ranking ones. It is left to the decision maker, when the ranking list of the alternatives is returned.

5. AHP and thesauri quality assessment

This section illustrates how to apply AHP to the thesaurus quality assessment, by considering three perspectives: TF creation, TF maintenance and TF indexing. The input data for our case study are the set of 24 vocabularies and the 21 quality criteria provided by the reference scenario. The performance values of Suominen and Mader [14] are used (Table 4), and different priorities are given to criteria according to the specific peculiarities of each context. Applying the AHP to these available data instead of computed performance values for the current thesauri releases does not affect the outcomes of our testing. Our aim is to provide a proof of concept of our methodology, rather than to supply an up to date assessment of the thesauri quality.

In the following subsections we introduce the peculiarities of the three application contexts, and describe the adoption of the AHP approach step by step, also analysing the results.

5.1. Context setting

Table 2 reports the assignment of priorities to the quality categories and issues, according to the three roles played by the decision makers: the thesaurus developer, TF manager and the indexing manager. We identified the priorities for AC1 and AC2 according to our experience in the deployment of the LusTRE framework with the consensus of the thesaurus developer of EARTh. The AC3 priority assignment is based on the issues affecting the indexing as reported by Haslhofer et al. [63]. We assumed three values for the criteria and category relevance: R (Relevant), L (Less Relevant), I (Irrelevant). Irrelevant criteria are not included in the hierarchy.

As regards context AC1, the quality assessment evaluates the appropriateness of including a thesaurus in the TF. The assignment of priorities is: (i) the ‘1 Authoritative Concepts’ category is relevant as the “Not too few / Not too many” rule of thumb has been defined to underestimate thesauri which, due to their dimensional characteristics, might turn out to be too poor or poorly maintained; (ii) the issues that can be fixed automatically by Skosify (indicated with an asterisk in Table 2) are judged as irrelevant. We could have considered such issues in AC1, but we prefer to fix these issues once the thesauri are added to the TF. We consider that selection and maintenance are two separate phases, and that new errors could be introduced by interlinking materialization after the thesauri are uploaded into the framework; (iii) issues such as ‘2.2 Incomplete language coverage’, ‘2.4 Overlapping labels’ and ‘3.1 Orphan concepts’, are particularly relevant for the effective reusability of the thesaurus by the end user, while the importance of other issues such as ‘3.4 Valueless associative relations’ and ‘3.6 Omitted top concepts’, are relaxed according to the needs of the decision maker. For instance ‘3.4 Valueless associative relations’ may be judged less relevant because skos:related are sometimes specialized with other properties, hence the fact that two siblings have an association might not necessary imply a bad modelling; (iv) the ‘4 Linked Data’ category has a lower relevance than the other three categories, as AC1 focuses on selecting thesauri to be inserted into the framework, and this insertion includes dedicated activities to improve interlinking.

As regards context AC2, the TQA focuses on the TF maintenance concerning the identification of thesauri (and their interlinking) deployed within the TF which may require fine-tuning with automatic procedures. Priorities are assigned as follows: (i) the ‘1 Authoritative Concepts’ category is not considered, because all the thesauri evaluated are assumed to be already included in the TF; (ii) the three other categories are significant: in particular issues that can be fixed automatically by Skosify (marked with an asterisk) are all relevant, while issues that intrinsically qualify the content of a thesaurus (such as ‘2.2 Incomplete language coverage’, ‘2.3 Undocumented concepts’ and ‘2.4 Overlapping labels’) are not considered at all in the TF maintenance. This is because they consist of editorial changes which a framework manager is not usually in charge of; (iii) all the issues in the ‘4 Linked Data’ category are relevant as the TF maintenance involves improving the interlinking between vocabularies in the TF and with other thesauri available as LD.

As regards context AC3, the TQA is aimed at identifying thesauri for the indexer/query expansion process. This context refers to indexing and query expansion as described in Haslhofer et al. [63], in which a weight is assigned to each SKOS property when computing the query expansion. The assignment of priorities is: (i) the “1 Authoritative Concepts” category is less relevant as such concepts can partially influence the richness of indexing: too few concepts might mean that the document will be poorly indexed, too many can also be problematic; (ii) the issues that can be solved automatically (i.e. with Skosify) are not considered as they should be solved before applying the indexing; (iii) "2 Labelling and Documentation" is particularly relevant: "2.2 Incomplete language coverage" is relevant because if a preferred label in a specific language is missing, that concept cannot be indexed in that language, "2.4 Overlapping labels" is relevant as overlapping labels can make indexing harder; (iv) "3 Structural" is less relevant as only a few of the issues listed in this category affect the indexing and query expansion. "3.1 Orphan concepts" and "3.2 Disconnected concept clusters" are relevant as they deal with single concepts or groups of concepts that are isolated; (v) the "4 Linked Data" category is relevant as interlinking between thesauri enlarges the set of terms that can be exploited in query expansion.

Table 2. Thesauri quality issues grouped by category and assignment of criteria priorities by the decision maker with respect to the application contexts AC1, AC2, AC3: R (Relevant), L (Less Relevant), I (Irrelevant).

Quality Issues Category	Quality issue	AC1	AC2	AC3
1 Authoritative Concepts		R	I	L
	1.1 Authoritative Concepts	R	I	R
2 Labelling and Documentation		R	R	R
	2.1 Omitted or invalid language tags (*)	I	R	I
	2.2 Incomplete language coverage	R	I	R
	2.3 Undocumented concepts	L	I	I
	2.4 Overlapping labels	R	I	R
	2.5 Inconsistent preferred labels (*)	I	R	I
	2.6 Disjoint labels violation (*)	I	R	I
	2.7 Extra whitespace in labels (*)	I	R	I
3 Structural		R	R	L
	3.1 Orphan concepts	R	I	R
	3.2 Disconnected concept clusters	R	I	R
	3.3 Cyclic hierarchical relations (*)	I	R	I
	3.4 Valueless associative relations	L	I	I
	3.5 Solely transitively related concepts (*)	I	R	I
	3.6 Omitted top concepts	L	L	I
	3.7 Top concepts having broader concepts	I	R	I
	3.8 Unidirectionally related concepts (*)	I	R	I
	3.9 Relation clashes (*)	I	R	I
	3.10 Mapping clashes	R	R	I
4 Linked Data		L	R	R
	4.1 Missing In-links	I	R	R
	4.2 Missing Out-links	L	R	R
	4.3 Broken links	R	R	R

* Issues fixable with Skosify.

5.2. Applying AHP

We now show how to apply the AHP phases introduced in Section 4.2. Several AHP tools are available [58], some of which require a fee, while others allow free usage or trials. We used SuperDecisions (http://www.superdecisions.com/) as we were granted a one-year free trial.

5.2.1. AHP Phase 1

With respect to the generic TQA scenario discussed in Section 4.2.1, only the criteria relevant for each specific application context are selected. Figure 2 illustrates the tree related to AC1, according to the priorities of quality criteria as reported in Table 2. If an issue is not relevant for the application context, there is no corresponding node in the tree. The tree root is labelled 'TQA' in order to recall the goal of the decision process. AC2 and AC3 trees are constructed similarly.

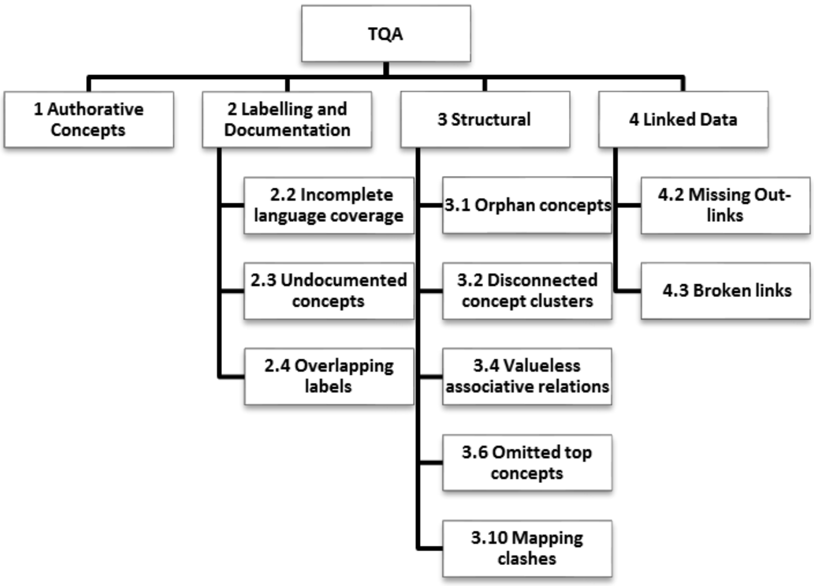


Figure 2. Phase 1 of AHP applied to AC1.

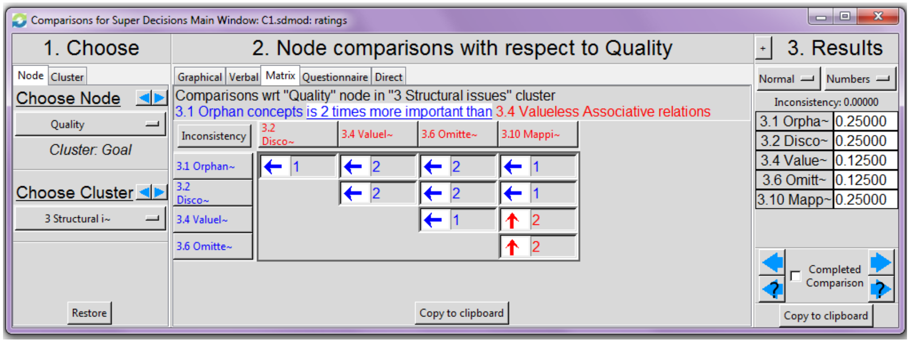
5.2.2. AHP Phase 2

Table 3 shows the pairwise comparison for the second level of the AC1 tree. The pairwise assessment of the four categories is given by the 4 x 4 matrix, where it is assumed that '1 Authoritative Concepts', '2 Labelling and Documentation' and '3 Structural' categories are, in this case, each individually twice as important as '4 Linked data' (see Table 2).

Table 3. Pairwise assessment of the four quality categories under the TQA goal with respect to AC1.

	1Authoritative Concepts	2 Labelling and Documentation	3 Structural	4 Linked Data
1 Authoritative Concepts	1	1	1	2
2 Labelling and Documentation	1	1	1	2
3 Structural	1	1	1	2
4 Linked Data	1/2	1/2	1/2	1

The same process is repeated for all the branches of all levels. Thus for AC1 we completed a 3x3 matrix, a 5x5 and a 2x2 matrix for the criteria at the leaves level. For instance, Figure 3 shows the elicitation of criteria priorities for the category '3 Structural' in AC1 with the use of SuperDecisions. According to the priorities, as presented in Table 2, SuperDecisions establishes the pairwise priorities amongst the five criteria belonging to '3 Structural'. The tool provides the insertion of priorities in a matrix form, however, as the inverse values are directly computed, just the triangular matrix needs to be filled. The priority of the last criterion (i.e. '3.10 Mapping clashes') compared to the others is not explicitly given, as it can be computed by the values supplied in the other cells. The weights are automatically computed by SuperDecisions according to Saaty [64]. The same process was applied to AC2 and AC3.



Comparisons wrt "Quality" node in "3 Structural issues" cluster
3.1 Orphan concepts is 2 times more important than 3.4 Valueless Associative relations

Inconsistency	3.2 Disco~	3.4 Valuel~	3.6 Omitte~	3.10 Mappi~
3.1 Orphan~	← 1	← 2	← 2	← 1
3.2 Disco~		← 2	← 2	← 1
3.4 Valuel~			← 1	↑ 2
3.6 Omitte~				↑ 2

Inconsistency: 0.00000

3.1 Orpha~	0.25000
3.2 Disco~	0.25000
3.4 Value~	0.12500
3.6 Omitt~	0.12500
3.10 Mapp~	0.25000

Figure 3. Pairwise comparison and computation of relative weights given in (triangular) matrix form for the Structural category according to the AC1 context. The normalized weights computed by the tool are shown on the right.

5.2.3. AHP Phase 3

The third AHP phase entails the computation of performance values of each criterion $\mathcal{C}_j$ according to each thesaurus. We only assigned values, on a subjective basis, to '1 Authoritative Concepts', and for the other 20 criteria we borrowed the performance values presented by Suominen and Mader [14]. Table 4 (in the Appendices) reports the number or errors for each issue before the correction by Skosify. We mapped values to the '1.1 Authoritative Concepts' indicator according to the "Not too few / Not too many" principle, as explained in Section 3.1. These values were assigned after categorizing the number of authoritative concepts based on the knowledge of the thesaurus expert. We used SuperDecisions to translate the expert knowledge in modelling the "Not too few / Not too many" rule of thumb. Thus five qualitative values were established: "Very Good", "Good", "Average", "Sufficient" and "Unsatisfactory". The mapping of thesauri into the category is subjective and it was also driven by our experience of some of them. For instance, DBpedia was set as "Unsatisfactory", since it provides about 850,000 authoritative concepts which present a variable level of accuracy since they are the result of crowdsourcing. ODT, GeoNames, and Reegle were also considered "Unsatisfactory" since they provide a limited number of concepts (below 1500); PXV NYTL, SSW, IPTC, UNESCO, Plant DDC were considered "Sufficient" (as they are between 1600 and 3500 or between 200000 and 400000 concepts); IOSV, NYTP, GEMET, STW, EuroVoc, SNOUMED, GTAA were considered as "Average" as they ranged between 3500 and 7000 concepts or between 35000 and 175000); LVAk and EARTh were considered as "Good" (between 7000 and 15000 concepts) and UMBEL, AGROVOC were considered as very good (between 15000 and 35000 concepts). SuperDecisions provides an easy interface to create this set of nominal categories, and to compare their relevance through the same pair-wise approach supporting the criteria comparison. Based on the user choices, the tool assigns a numerical value to each qualitative value which, in our specific case, are 1, 0.62, 0.369, 0.211, 0.106, respectively (normalized with respect to the highest value).

5.2.4. AHP phase 4

Before applying the fourth phase of the AHP methodology, i.e. the evaluation of the scores $\mathcal{S}_{ij}$ for each criterion $\mathcal{C}_j$ for each thesaurus $\mathcal{T}_j$ (i=1,...,24), a pre-processing step is needed to enable SuperDecisions to integrate the performance quality values. This pre-processing is usually not necessary in decision scenarios that adopt AHP to compare alternatives, for instance when evaluating the price of goods, the air pollution caused by alternative producing plants, the horse power and energy consumption of vehicles and so on. However, in our case the performance values were provided as the absolute number of errors for any given IQ issue and thesaurus, while the assessed thesauri have different sizes (i.e. different numbers of authoritative concepts). Thus prior to being normalized, such performance values need to be measured on a comparable scale i.e. their absolute values need to be relativized to the thesaurus size.

We therefore replaced the absolute performance value $\mathcal{P}_{ij}$ for a criterion $\mathcal{C}_j$ given an alternative $\mathcal{T}_i$, with the ratio:

$$\mathcal{P}_{ij} = \frac{\#\text{error}_{ij}}{\#\text{authconcept}(\text{T}_i)} \tag{1}$$

In (1) with the authoritative concepts as the denominator, $\mathcal{P}_{ij}$ represents the average errors for a concept given criterion C_j, which is a homogenous and natural indicator of the lack of quality among thesauri of different sizes. Once all ratios have been computed for each criterion and alternative, the usual AHP scaling phase can be applied. Considering that the best result is when $\mathcal{P}_{ij} = 0$ (i.e. no errors), we cannot directly use the Ideal approach, mentioned in Section 4.2.4, which normalizes the values for a given criterion by dividing by the maximum (assuming that this is the best one). We therefore applied the following normalization formula to compute the score value S_{ij}:

$$S_{ij} = \begin{cases} 1 - \frac{\mathrm{P}_{ij}}{max_j}, & max_j \neq 0 \\ 1, & max_j = 0 \end{cases} \tag{2}$$

where max$_j$ is the maximum (relative) error measured for a given criterion C_j(i.e. max$_j$ =max $\{\mathcal{P}_{ij}$, i=1,..,24$\}$). Thus, as expected, the smaller the error, the greater the score achieved by a thesaurus for a given criterion, which then increases the ranking of the thesaurus when coupled with the computed weights. If for a criterion C_k, no error is reported for any thesaurus $\mathcal{T}_i$, then $S_{ik} = 1$ for all i. Table 5 in the Appendices reports the computed scores, on the basis of the absolute performance values contained in Tables 6, 7 and 8 in [14] which are summarized in Table 4 of the present paper.

5.2.5. AHP Phase 5

Based on the overall synthesis score achieved by each thesaurus, following the process explained in Section 4.2.5 SuperDecisions computes and returns a ranking list of the 24 alternative thesauri. The furthest three columns on the right of Table 5 show the overall synthesis scores, and Figure 4 summarizes the ranking for the three contexts.

5.3. Results

Figure 4 highlights that different application contexts result in different thesauri rankings: 17 out of the 25 thesauri were ranked at different positions for the three different contexts. For instance, GeoNames ranks 23rd, 5th and 18th respectively in AC1, AC2 and AC3. Rankings are based on the overall thesauri scores, computed by aggregating the criteria scores according to the synthesis phase explained in Section 4.2.5. Breaking down the overall score into its subcomponents highlights the contribution of each issue category in ranking each thesaurus. This could provide decision makers with an overview of the impact of their judgments on criteria priorities.

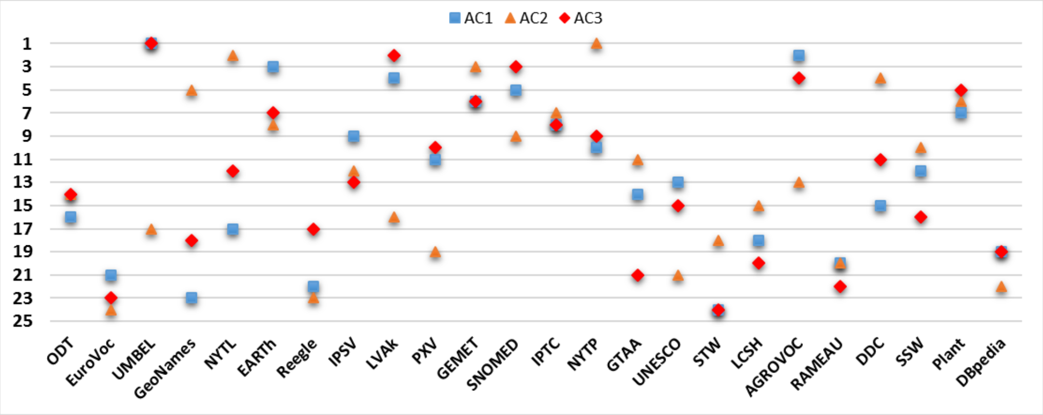


Figure 4. Ranking of the 24 thesauri according to contexts AC1, AC2 and AC3.

Let us consider, for instance, UMBEL which ranks first in AC1 but 17th in AC2, and NYTP which ranks 10th in AC1 and first in AC2. The overall thesaurus qualities with respect to AC1 are 0.969 and 0.701, respectively (see Table 5). We show the behaviour of the two thesauri according to the partial scores based on the criteria priorities provided by the decision maker. For the ‘1.1 Authoritative Concepts’ criterion, UMBEL scores 0.286=1**0.286* versus NYTP scoring 0.105=0.369**0.286* (hereafter we mark the weights computed by the SuperDecisions tool in italics). For the ‘2 Labelling and Documentation’ category, UMBEL scores 0.264=(1**0.4*+0.892**0.2*+0.868**0.4*)**0.286* versus NYTP scoring 0.239=(1**0.4*+0.178**0.2*+1**0.4*)**0.286*. For the ‘3 Structural’ category UMBEL scores 0.276=(0.889**0.25*+0.973**0.25*+1**0.125*+1**0.125*+1**0.25*)**0.286* versus NYTP scoring 0.214=(0**0.25*+1**0.25*+1**0.125*+0.985**0.125*+1**0.25*)**0.286*. For the ‘4 Linked Data’ category, UMBEL and NYPT score the same partial value of 0.143=(1**0.3333*+1**0.6667*)**0.143*. From the decision maker’s point of view, these partial scores highlight that: UMBEL noticeably outperforms NYTP in terms of the first category, is better than NYTP regarding the labelling and documentation issues and the structural issues, and is equal to NYTP in terms of linked data issues.

The overall thesaurus qualities with respect to AC2 are 0.750 for UMBEL and 0.893 for NYTP. Without repeating the detailed analysis of each category’s partial score, it is reasonable that such differences are largely affected by the lack of relevance given to the issue ‘1.1 Authoritative Concepts’ in the AC2 context as explained in Section 5.1.

To further evaluate the results of the empirical analysis, a repeated measure analysis of variance (ANOVA) was performed to explore the differences of each application context. Descriptive statistics for the overall quality measures yielded the following values: AC1 (Mean: 0.691, SD: 0.102), AC2 (Mean: 0.777, SD: 0.070), AC3 (Mean: 0.670, SD: 0.096). The means of the overall quality scores for the different contexts were significantly different: F(2.69)= 9.491, p=0.0002. A p-value lower than 0.01 strongly suggests that one or more pairs of treatments are significantly different. We therefore applied Tukey's HSD test to each of the three treatment pairs to pinpoint which one exhibits statistically significant differences. The results show that there are statistically significant differences between AC1 and AC2 (p=0.004) and between AC2 and AC3 (p=0.001), while there is no statistically significant difference between AC1 and AC3 (p= 0.689).

6. Discussion

There are many real case scenarios in which AHP has already been successfully adopted, although have been some criticism over the years, which were countered by Saaty himself and other methodology proponents [65]. In our approach we highlight two main AHP properties. Firstly, AHP supports the assessment of the trade-off between criteria at the various levels of the hierarchy. The hierarchy organization supports the grouping of issues into categories of the TQA described in Section 3.1. Secondly, AHP facilitates the thesauri quality comparison for decision makers. Indeed, as highlighted by Saaty [66]:

> Comparison requires the use of judgment. Even informed judgment is subjective.

Conversely, if no categorization of issues and the associated hierarchic judgement modelling are needed, a simpler and more widely known approach such as Simple Additive Weighting (SAW), could be devised. With SAW, decision makers directly assign a weight w_j to each criteria C_j, where $\sum w_j$=1.

In order to appraise the impact of criteria priorities on the final ranking, we apply SAW by assuming all criteria as equally relevant, i.e. no context is defined. Considering the testbed thesauri T_i, (i=1,..,24), and all 21 criteria as equally relevant, i.e. w_j = 1/21, SAW aggregates the 21 criteria scores S_{ij}according to the sum:

$$S_i = \sum_{j=1}^{21} w_j S_{ij} \tag{3}$$

The overall scores S_i are thus only affected by the criteria performance scores. Figure 5 shows the resulting thesauri ranking with those related to AC1, AC2 and AC3. The equality of weights applied to the SAW scenario altered the relative ranks for the majority of the thesauri. There are four exceptions (i.e. ODT, Rameau, Plant, and DBpedia) with a limited ranking variation (i.e. ±2) for all the contexts. This implies that none of the rankings obtained in Figure 4 would have been obtained if we had not explicitly considered the contexts, and confirms the need for context modelling according to the decision makers’ needs.

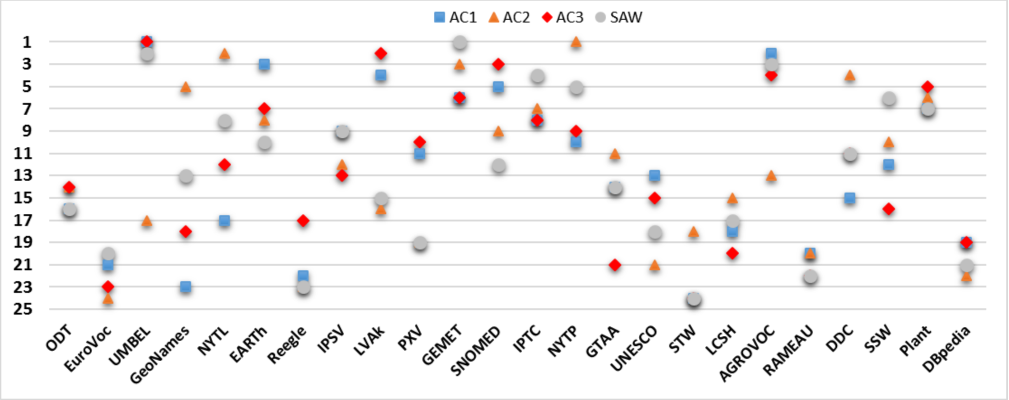


Figure 5. Ranking variations with respect to AC1, AC2, AC3 and SAW method applied with all criteria being equally relevant.

Considering how the ranking of alternatives is affected by subjective judgments, a common practice in MCDM applications is to adjust the set of criteria, or their relative importance. This leads to a feedback process, namely sensitive analysis that supports decision makers in testing the influence of different criteria settings on the alternative rankings. This analysis is useful both in cases of doubt regarding the ranking of alternatives, and when studying the impact of different priorities. The analysis can take place for instance after the identification of new relevant criteria or when changing the usage scenario (i.e. different expectations on the set of criteria) [67,68].

Figure 6 reports the ranking of thesauri after modifying the priorities of criteria for the context AC1. The modified AC1, namely AC1', considers relevant the set of criteria fixable by Skosify which were judged irrelevant in AC1. As explained in Section 5.1, this is an alternative description of the thesauri selection operated by a thesaurus developer. AC1', does not substantially subvert the ranking. For instance, the first three thesauri (i.e. UMBEL, AGROVOC and EARTh) are still the same, and there are very limited variations for the other ranks: for 17 thesauri there is a ranking variation of at most ±3 positions and the greatest differences are 6 positions for PXV and 5 positions for DDC and NYTP. From a thesaurus developer's point of view, the criteria modifications are not disruptive, and the original modelling of AC1 is feasible.

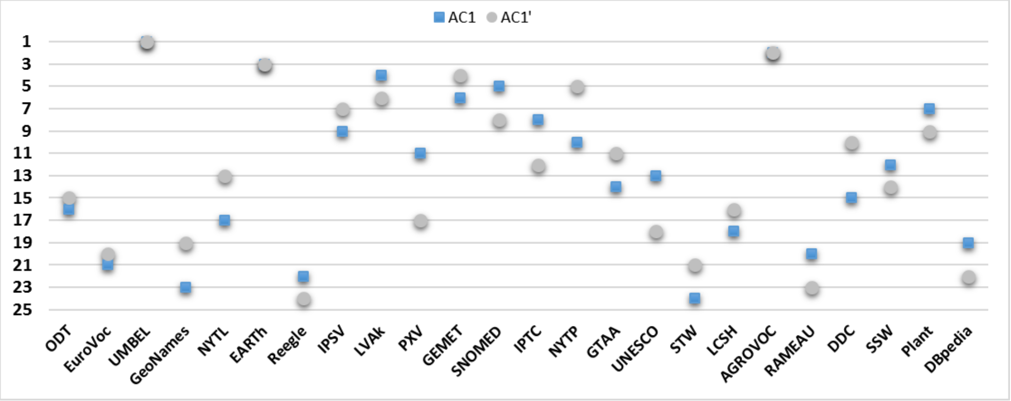


Figure 6. Ranking variations with respect to AC1 when a sensitive analysis has altered criteria priorities (AC1').

7. Conclusions

We have proposed the adoption of the AHP as a means to assess the quality of SKOS thesauri. Unlike other proposals, which define metrics by measuring distinct pieces of linked datasets, we focus on providing an overall quality measure that synthesizes a heterogeneous set of quality dimensions. As a guideline to replicate our approach, we have provided a step-by-step explanation of how AHP supports the decision tasks involved in three application contexts: the creation, maintenance and exploitation of linked thesauri. The analysis of the results highlights the appropriateness of modelling different application contexts in order to meet decision makers' needs in thesauri selection. AHP is able to capture both subjective and objective facets involved in the selection of thesauri, and to provide a ranking of the assessed thesauri.

Ongoing research is extending the use of the AHP methodology to an overall quality assessment of thesauri available in the linked open data cloud by extending the reference scenario to additional quality issues such as those of linksets and quality related to the compliance with W3C best practices.

Funding

This research was partially carried out within the EU-funded project eENVplus (CIP-ICT-PSP grant No. 325232).

References

[1] Hodge G. 'Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files', Council on Library and Information Resources Publications, http://www.clir.org/pubs/reports/reports/pub91/pub91.pdf (2000, accessed June 2016).
[2] Shiri AA, Revie C and Chowdhury G. Thesaurus-enhanced search interfaces. Journal of Information Science 2002; 28(2): 111-22.
[3] Albertoni R, De Martino M, Di Franco S, De Santis V and Plini P. EARTh: An Environmental Application Reference Thesaurus in the Linked Open Data cloud. Semantic Web journal 2014; 5(2): 165–171.
[4] Caracciolo C, Stellato A, Morshed A, et al. The AGROVOC linked dataset. Semantic Web journal 2012; 4(2): 341–348.
[5] Zapilko B, Schaible J, Mayr P and Mathiak B. TheSoz: A SKOS representation of the thesaurus for the social sciences. Semantic Web journal 2013; 4(3): 257–263.
[6] Miles A and Bechhofer S. 'W3C Recommendation: Simple Knowledge Organization System Reference', http://www.w3.org/TR/skos-reference (2009, accessed June 2016).
[7] Berners-Lee T: Linked data, http://www.w3.org/DesignIssues/LinkedData.html (2009, accessed June 2016).
[8] Heath T and Bizer C. Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool 2011.
[9] De Martino M and Albertoni R. A multilingual/multicultural semantic-based approach to improve Data Sharing in a SDI for Nature Conservation. International Journal of Spatial Data Infrastructures Research 2011; 6: 206-233.
[10] Abecker A, Albertoni R, De Martino M, Podestà P, Schnitter K and Wössner R. Latest Developments of the Linked Thesaurus Framework for the Environment (LusTRE), In: Proceedings of the 29th EnviroInfo 2015 Conference, Copenhagen, 2015.
[11] Palavitsinis N and Manouselis N. A Survey of Knowledge Organization Systems in Environmental Sciences. In: Proceedings of the 4th International ICSC Symposium, ITEE 2009, Thessaloniki, 505–517.
[12] Albertoni R, De Martino M and Podestà P. Environmental Thesauri under the Lens of Reusability. In EGOVIS 2014 Conference, Munich, September 2014, LNCS 8650: 222-236.
[13] Zaveri A, Rula A, Maurino A, Pietrobon R, Lehmann J and Auer S. Quality assessment for linked open data: A survey. Semantic Web journal 2016; 7(1): 63-93.
[14] Suominen O and Mader C. Assessing and Improving the Quality of SKOS Vocabularies. Journal on Data Semantics 2013; 3: 47–73.
[15] Saaty TL. The Analytic Hierarchy Process. New York: McGraw-Hill, 1980.
[16] Bawden D and Robinson L. The dark side of information overload, anxiety and other paradoxes and pathologies. Journal of Information Science 2009; 35(2): 180-191.
[17] Ouzzani M, Papotti P and Rahm E. Introduction to the special issue on data quality. Information Systems 2013; 38: 885–886.
[18] Wang RY and Strong, D. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems 1996; 12(4): 5–34.
[19] Tayi GK and Ballou DP. Examining data quality. Communications of the ACM 1998; 41(2): 54-57.
[20] Bordogna G, Carrara P, Criscuolo L, Pepe M and Rampini A. A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science. Information Sciences 2014; 258: 312-327.
[21] Juran JM. Quality Control Handbook. 3rd ed. New York: McGraw-Hill Book Co, 1974.
[22] Bizer C and Cyganiak R. Quality-driven information filtering using the WIQA policy framework. Journal of Web Semantics 2009; 7(1):1–10.
[23] Naumann F. Quality-driven Query Answering for Integrated Information Systems. Lecture Notes in Computer Science 2002; 2261: 1-166.
[24] Redman TC. Data Quality for the Information Age. London: Artech House, 1997.

[25] Delone WH and McLean ER. Information systems success: the quest for the dependent variable. Information Systems Research 1992; 3(1):60–95.

[26] Pirsig R. Zen and the Art of Motocycle Maintenance. New York: Bantam Books, 1974.

[27] Batini C, Cappiello C, Francalanci C and Maurino A. Methodologies for data quality assessment and improvement. ACM Computing Surveys 2009; 41(3): 16:1-16:52.

[28] Kim KS and Sin SCJ. Selecting quality sources: Bridging the gap between the perception and use of information sources. Journal of Information Science 2011; 37(2): 178–188.

[29] Naumann F. ‘Information Quality: Fundamentals, Techniques, and Use’, EDBT Tutorial, http://www.hpi.uni-potsdam.de/fileadmin/hpi/FG_Naumann/publications/EDBT06Tutorial_IQ.pdf (2006, accessed June 2016).

[30] DeMarco T. Controlling software projects: management measurement and estimation. New York: Yourdon, 1982.

[31] Guéret C, Groth P, Stadler C and Lehmann J. Assessing linked data mappings using network measures. In: ESWC 2012, LNCS 7295, 87–102.

[32] Albertoni R and Gómez-Pérez A. Assessing linkset quality for complementing third-party datasets. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops, ACM, New York 2013, 52-59.

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

[34] Demartini G, Difallah DE and Cudré-Mauroux P. Large-scale linked data integration using probabilistic reasoning and crowdsourcing; The International Journal on Very Large Data Bases 2013; 22(5): 665-687.

[35] Debattista J, Lange C and Sören A. ‘LUZZU – A Framework for Linked Data Quality Assessment’, CoRR abs/1412.3750, http://arxiv.org/abs/1412.3750 (2015, accessed June 2016).

[36] Kless D and Milton S. Towards quality measures for evaluating thesauri. In: Proceedings of the 4th metadata and semantics research conference (MTSR 2010), Communications in computer and information science 2010; 108: 312–319.

[37] Albertoni R, De Martino M, and Podestà P. A Linkset Quality Metric Measuring Multilingual Gain in SKOS Thesauri. In: Proceedings of 2nd Workshop on Linked Data Quality co-located with 12th Extended Semantic Web Conference (ESWC 2015).

[38] Sabbah T, Selamat A, Ashraf M and Herawan T. Effect of thesaurus size on schema matching quality. Knowledge-Based Systems 2014; 71: 211–226.

[39] Mader C, Haslhofer B and Isaac A. Finding quality issues in SKOS vocabularies. In: Proceedings of the second international conference on theory and practice of digital libraries (TPDL 2012). Lecture Notes in Computer Science 2012; 7489: 222–233.

[40] Suominen O and Hyvönen E. Improving the quality of SKOS vocabularies with Skosify. In: Proceedings of the 18th international conference on knowledge engineering and knowledge management, (EKAW 2012). Lecture Notes in Computer Science 2012, 7603: 383–397.

[41] Pipino L, Lee Y and Wang R. Data quality assessment. Communication ACM 2002; 45(4): 211–218.

[42] Jiménez Martín A, Suárez-Figueroa MC, Mateos Caballero A, Gómez-Pérez A, Fernández-López M. A Maut aprroach for reusing domain ontologies on the basis of the NeOn Methodlogy. International Journal of Information Technology & Decision Making (IJITDM) 2013; 12(5): 945-968.

[43] Lozano-Tello A and Gómez-Pérez A. ONTOMETRIC: A Method to Choose the Appropriate Ontology. Journal of Database Management 2004; 2 (15): pp. 1-18.

[44] Guarino N and Welty C. Evaluating ontological decisions with OntoClean. Communications of the ACM 2002; 45(2): 61-65.

[45] Hwang SH, Kim HG, Yang HS. A FCA-Based Ontology Construction for the Design of Class Hierarchy. Computational Science and Its Applications–ICCSA 2005; 2005: 307-20.

[46] Working Group on Guidelines for Multilingual Thesauri, ‘Guidelines for Multilingual Thesauri’, IFLA Professional Reports 115, http://www.ifla.org/files/assets/hq/publications/professional-report/115.pdf (2009, accessed June 2016).

[47] Bergamin G and Lucarelli A. The Nuovo soggettario as a service for the linked data world. Italian Journal of Library, Archives, and Information Science 2013; 4(1): 213-226.

[48] Suominen O, Pessala S, Tuominen J, Lappalainen M, Nykyri S, Ylikotila H, et al. Deploying National Ontology Services: From ONKI to Finto. In Proceedings of the ISWC 2014 Industry track (2014).

[49] International Society on Multiple Criteria Decision Making, ‘Mission of the Society’, http://www.mcdmsociety.org/content/mission-society (accessed June 2016).

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

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

[52] Hwang CL and Yoon K. Multiple Attribute Decision Making: Methods and Applications. New York: Springer-Verlag, 1981.

[53] Roy B. Classement et choix en présence de points de vue multiples (la méthode ELECTRE). La Revue d'Informatique et de Recherche Opérationelle 1968; 8: 57–75.

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

[55] Ishizaka A, Pearman C and Nemery P. AHPSort: An AHP-based method for sorting problems. International Journal of Production Research 2012; 50(17): 4767–4784.

[56] Vaidyaa OS and Kumar S. Analytic hierarchy process: An overview of applications. European Journal of Operational Research 2006; 169(1): 1–29.

[57] Forman EH and Selly MA. 'Decision by Objectives', http://professorforman.com/DecisionByObjectives/DBO.pdf (accessed June 2016).

[58] Mustajoki J and Marttunen M. 'Comparison of Multi-Criteria Decision Analytical Software: Searching for ideas for developing a new EIA-specific multi-criteria software, IMPERIA Project Report', http://imperia.jyu.fi/tuotokset/Annex7.5.13ComparisonofMultiCriteriaDecisionAnalyticalSoftware.pdf (2013, accessed June 2016).

[59] Mesarovic MD and Macko D. Foundations for a scientific theory of hierarchical systems. Hierarchical Structures. New York: American Elsevier 1969, pp. 29-50.

[60] Saaty TL. A scaling method for priorities in hierarchical structures. Journal of Mathematical Psychology 1977; 15: 234-281.

[61] Saaty TL and Vargas LG. Comparison of eigenvalue, logarithmic least squares and least squares methods in estimating ratios. Mathematical Modeling 1984; 5: 309–324.

[62] Saaty TL and Vargas LG. The Seven Pillars of the Analytic Hierarchy Process. In Saaty TL, Vargas G Models, Methods, Concepts & Applications of the Analytic Hierarchy Process. International Series in Operations Research & Management Science 175. New York: Springer Science+Business Media, 2012, 23 – 40.

[63] Haslhofer B, Martins F and Magalhães J. Using SKOS vocabularies for improving web search. In Proceedings of the 22nd International Conference on World Wide Web (WWW '13 Companion) 2013. ACM, New York, NY, USA, 1253-1258.

[64] Saaty TL. The Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process. Pittsburgh: RWS Publications, 2011.

[65] von Solms SH. Validity of the AHP/ANP: Comparing Apples and Oranges. International Journal of the Analytic Hierarchy Process 2011; 3(1): 2-27.

[66] Saaty, TL. Reflections and Projections on Creativity in Operations Research and Management Science: A Pressing Need for a Shift in Paradigm. Operations Research 1998; 46(1): 9-16.

[67] Triantaphyllou E, and Sanchez A. A Sensitivity Analysis Approach for Some Deterministic Multi-Criteria Decision-Making Methods. Decision Sciences 1997; 28(1): 151-194.

[68] Masuda T. Hierarchical sensitivity analysis of the priorities used in the Analytic Hierarchy Process, Systems Science 1990; 21(2): 415-427.

Appendices

Table 4. Number of errors for quality issues (as reported in tables 6, 7 and 8 of Suominen and Mader [14]).

	1 A. C.	2 Labelling and Documentation							3 Structural										4 Linked Data		
	1.1 Authoritative concepts	2.1 Omitted or invalid language tags (*)	2.2 Incomplete language coverage	2.3 Undocumented concepts	2.4 Overlapping labels	2.5 Inconsistent preferred labels (*)	2.6 Disjoint labels violation (*)	2.7 Extra whitespace in labels (*)	3.1 Orphan concepts	3.2 Disconnected concept clusters	3.3 Cyclic hierarchical relations (*)	3.4 Valueless associative relations	3.5 Solely transitively related concepts (*)	3.6 Omitted top concepts (*)	3.7 Top concepts having broader concepts	3.8 Unidirectionally related concepts (*)	3.9 Relation clashes (*)	3.10 Mapping clashes	4.1 Missing In-links	4.2 Missing Out-links	4.3 Broken links
ODT	107	3	16	35	2	0	1	0	4	7	0	7	0	0	2	126	0	0	111	31	37
EuroVoc	6797	219	6370	5341	62	0	0	2	7	4	0	6	0	1	0	14289	0	0	6170	6797	120790
UMBEL	26389	25793	0	2848	5207	2	1	522	2936	86	5	0	36535	0	0	740	0	0	26110	0	130
GeoNames	680	0	43	60	162	1	0	0	680	0	0	0	0	9	0	0	0	0	24	680	11
NYTL	1920	0	0	1862	0	0	0	0	1920	0	0	0	0	1	0	0	0	0	1892	0	1376
EARTh	14351	10	313	7840	2100	0	69	310	2288	354	0	1124	0	0	0	12091	61	0	14349	9558	410
Reegle	1447	3	1450	3	22	0	3	52	4	2	0	2013	842	0	0	1718	317	2	1447	809	321
IPSV	4732	0	0	4551	0	0	21	0	0	1	0	253	0	0	0	25	5	0	4731	4732	1
LVAk	13411	13411	0	13411	13	0	0	0	21	11	5	5	0	0	0	16344	1	0		13411	
PXV	1686	1578	0	1492	7	0	4	2	2	10	0	0	0	0	1	2725	2	0	1686	1046	107
GEMET	5209	4	894	1	3638	0	3	12	0	5	0	31	0	1	0	9657	2	0	3290	584	40
SNOMED	102614	102600	0	102614	229	0	202	0	0	1	0	119	0	0	0	60396	1234	0	102610	0	5
IPTC	2061	0	0	933	1	0	0	0	0	10	0	0	1113	0	0	2241	0	0	2061	933	2
NYTP	4979	0	0	4094	0	0	0	6	4979	0	0	0	0	1	0	0	0	0	4965	0	9
GTAA	171991	0	0	96850	11894	0	0	0	162000	621	0	9448	0	9	0	18804	37	0	171990	171991	740
UNESCO	2509	0	0	2509	227	0	0	1524	0	1	0	19	0	0	0	124	0	0	2509	2509	1
STW	6789	47	25050	5290	10123	214	0	0	70	141	0	5004	0	2	0	18533	5	0	6781	1463	504
LCSH	408923	100317	0	308607	7766	669	206	0	173149	22343	0	0	0	1	0	96533	0	0	408920	347560	2640
AGROVOC	32291	0	32060	29820	2666	0	2424	2166	0	234	0	281	0	0	0	20672	1	0	31680	17286	160
RAMEAU	207272	116343	140860	70358	5539	0	33066	7940	86137	24927	4	5118	0	0	0	322079	377	0	207260	34803	132333
DDC	251977	0	158161	251977	40729	1	0	416	97294	2087	0	0	0	30	1812	4761	0	0	250790	458	110
SSW	1943	4	1143	1328	39	0	16	6	6	1	0	118	22	0	0	723	4	0	1941	1606	285
Plant	3246	1	0	220	54	0	0	0	0	22	0	3463	0	0	44	3246	0	0	3246	0	662
DBpedia	865902	0	0	865902	765	0	0	0	103877	1174	1133	9021	0	0	0	1713339	10219	0	865566	865902	11400

* Issues fixable with Skosify.

Table 5. Normalized performance values of quality criteria and overall quality computed for the three application contexts AC1, AC2 and AC3.

	1 A. C.	2 Labelling and Documentation							3 Structural										4 Linked Data			Overall quality		
	1.1 Authoritative concepts	2.1 Omitted or invalid language tags (*)	2.2 Incomplete language coverage	2.3 Undocumented concepts	2.4 Overlapping labels	2.5 Inconsistent preferred labels (*)	2.6 Disjoint labels violation (*)	2.7 Extra whitespace in labels (*)	3.1 Orphan concepts	3.2 Disconnected concept clusters	3.3 Cyclic hierarchical relations (*)	3.4 Valueless associative relations	3.5 Solely transitively related concepts (*)	3.6 Omitted top concepts (*)	3.7 Top concepts having broader concepts	3.8 Unidirectionally related concepts (*)	3.9 Relation clashes (*)	3.10 Mapping clashes	4.1 Missing In-links	4.2 Missing Out-links	4.3 Broken links	**AC1**	**AC2**	**AC3**
ODT	0.106	0.972	0.959	0.673	0.987	1.000	0.941	1.000	0.963	0.456	1.000	0.953	1.000	1.000	0.000	0.569	1.000	1.000	0.000	0.710	0.981	0,661	0,774	0,648
EuroVoc	0.369	0.968	0.746	0.214	0.994	1.000	1.000	1.000	0.999	0.995	1.000	0.999	1.000	0.989	1.000	0.230	1.000	1.000	0.125	0.000	0.000	0,601	0,638	0,531
UMBEL	1.000	0.023	1.000	0.892	0.868	0.998	1.000	0.967	0.889	0.973	0.855	1.000	0.000	1.000	1.000	0.990	1.000	1.000	0.046	1.000	1.000	0,969	0,750	0,860
GeoNames	0.106	1.000	0.983	0.912	0.840	0.953	1.000	1.000	0.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	0.966	0.000	0.999	0,564	0,855	0,623
NYTL	0.211	1.000	1.000	0.030	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	0.961	1.000	1.000	1.000	1.000	0.050	1.000	0.960	0,643	0,889	0,675
EARTh	0.620	0.999	0.994	0.454	0.902	1.000	0.970	0.964	0.841	0.795	1.000	0.944	1.000	1.000	1.000	0.691	0.981	1.000	0.036	0.334	0.998	0,789	0,796	0,708
Reegle	0.106	0.998	0.728	0.998	0.990	1.000	0.987	0.941	0.997	0.989	1.000	0.000	0.580	1.000	1.000	0.565	0.000	0.000	0.036	0.441	0.988	0,576	0,677	0,632
IPSV	0.369	1.000	1.000	0.038	1.000	1.000	0.972	1.000	1.000	0.998	1.000	0.962	1.000	1.000	1.000	0.998	0.995	1.000	0.036	0.000	1.000	0,710	0,779	0,669
LVAk	0.620	0.000	1.000	0.000	0.999	1.000	1.000	1.000	0.998	0.993	0.715	1.000	1.000	1.000	1.000	0.554	1.000	1.000	1.000	0.000	1.000	0,786	0,768	0,825
PXV	0.211	0.064	1.000	0.115	0.997	1.000	0.985	0.998	0.999	0.951	1.000	1.000	1.000	1.000	0.968	0.408	0.995	1.000	0.036	0.380	0.996	0,690	0,712	0,687
GEMET	0.369	0.999	0.953	1.000	0.532	1.000	0.996	0.996	1.000	0.992	1.000	0.996	1.000	0.985	1.000	0.321	0.998	1.000	0.391	0.888	1.000	0,759	0,884	0,736
SNOMED	0.369	0.000	1.000	0.000	0.999	1.000	0.988	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.784	0.945	1.000	0.036	1.000	1.000	0,762	0,795	0,787
IPTC	0.211	1.000	1.000	0.547	1.000	1.000	1.000	1.000	1.000	0.960	1.000	1.000	0.610	1.000	1.000	0.602	1.000	1.000	0.036	0.547	1.000	0,724	0,802	0,708
NYTP	0.369	1.000	1.000	0.178	1.000	1.000	1.000	0.998	0.000	1.000	1.000	1.000	1.000	0.985	1.000	1.000	1.000	1.000	0.039	1.000	1.000	0,701	0,893	0,705
GTAA	0.369	1.000	1.000	0.437	0.954	1.000	1.000	1.000	0.058	0.970	1.000	0.961	1.000	0.996	1.000	0.960	0.999	1.000	0.036	0.000	1.000	0,664	0,780	0,588
UNESCO	0.211	1.000	1.000	0.000	0.939	1.000	1.000	0.000	1.000	0.997	1.000	0.995	1.000	1.000	1.000	0.982	1.000	1.000	0.036	0.000	1.000	0,662	0,698	0,640
STW	0.369	0.993	0.000	0.221	0.000	0.000	1.000	1.000	0.990	0.827	1.000	0.470	1.000	0.978	1.000	0.000	0.997	1.000	0.037	0.785	0.996	0,503	0,733	0,415
LCSH	0.211	0.755	1.000	0.245	0.987	0.948	0.997	1.000	0.577	0.546	1.000	1.000	1.000	1.000	1.000	0.914	1.000	1.000	0.036	0.150	1.000	0,627	0,769	0,592
AGROVOC	1.000	1.000	0.731	0.077	0.945	1.000	0.529	0.890	1.000	0.940	1.000	0.994	1.000	1.000	1.000	0.765	1.000	1.000	0.054	0.465	1.000	0,880	0,775	0,776
RAMEAU	0.211	0.439	0.816	0.661	0.982	1.000	0.000	0.937	0.584	0.000	0.985	0.982	1.000	1.000	1.000	0.431	0.992	1.000	0.036	0.832	0.964	0,619	0,705	0,587
DDC	0.211	1.000	0.830	0.000	0.892	1.000	1.000	0.997	0.614	0.931	1.000	1.000	1.000	0.991	0.615	0.993	1.000	1.000	0.041	0.998	1.000	0,653	0,873	0,677
SSW	0.211	0.998	0.841	0.317	0.987	1.000	0.948	0.995	0.997	0.996	1.000	0.956	0.992	1.000	1.000	0.864	0.991	1.000	0.037	0.173	0.992	0,674	0,787	0,639
Plant	0.211	1.000	1.000	0.932	0.989	1.000	1.000	1.000	1.000	0.944	1.000	0.233	1.000	1.000	0.275	0.634	1.000	1.000	0.036	1.000	0.989	0,737	0,836	0,754
DBpedia	0.106	1.000	1.000	0.000	0.999	1.000	1.000	1.000	0.880	0.989	0.000	0.993	1.000	1.000	1.000	0.275	0.946	1.000	0.036	0.000	0.999	0,630	0,691	0,622

* Issues fixable with Skosify.