"WHAT IS OUV" REVISITED: A COMPUTATIONAL INTERPRETATION ON THE STATEMENTS OF OUTSTANDING UNIVERSAL VALUE

Nan Bai[1], Pirouz Nourian[2], Renqian Luo[3], Ana Pereira Roders[1]

[1]UNESCO chair in Heritage and the Reshaping of Urban Conservation for Sustainability,
Chair of Heritage and Values, Delft University of Technology, Delft, the Netherlands
[2]Chair of Design Informatics, Delft University of Technology, Delft, the Netherlands
[3]University of Science and Technology of China, Hefei, China
{n.bai, p.nourian, a.r.pereira-roders}@tudelft.nl, lrq@mail.ustc.edu.cn

KEY WORDS: Outstanding Universal Value, UNESCO, World Heritage, Natural Language Processing, Machine Learning, Similarity Matrix, Graph Visualization, Statistics

ABSTRACT:

The Statements of Outstanding Universal Value (OUV) concerns the core justification for nominating and inscribing cultural and natural heritage properties on the UNESCO World Heritage List, ever since 2007. Ten criteria are specified and measured independently for the selection process. The 2008 ICOMOS Report "*What is OUV*" has been a successful example to interpret OUV as an integral concept by inspecting the associations of the selection criteria in all inscribed properties. This paper presents a novel methodology for interpreting OUV using computational techniques of Natural Language Processing, Machine Learning, and Graph Visualization. Firstly, frequent phrases appearing in Statements of OUV are used to construct a lexicon for each selection criterion; Secondly, three similarity matrices are constructed as graphs to represent the pair-wise associations of the criteria; Lastly, the lexicon and graphs are visualized in 2D. The study shows that the lexicon derived from computational techniques can capture the essential concepts of OUV, and that the selection criteria are consistently associated with each other in different similarity metrics. This study provides a quantitative and qualitative interpretation of the Statements of OUV and the associations of selection criteria, which can be seen as an elaborated computational extension of the 2008 Report, useful for future inscription and evaluation process of World Heritage nominations.

1. INTRODUCTION

The *World Heritage Convention* seeks to preserve the "*parts of the cultural and natural heritage ... of outstanding interest ... [for] mankind as a whole*" since its adoption in 1972 (UNESCO, 1972). A total of 1121 World Heritage (WH) properties have been inscribed on the World Heritage List until 2019. After the adoption of the *Operational Guidelines* in 2005, the justification of Outstanding Universal Value (OUV) has become an administrative requirement, instead of an independent qualification since 1977, for inscribing any new WH nomination (UNESCO, 2008, Jokilehto, 2008). Ten selection criteria exist as the core of OUV, among which criteria (i) - (vi) generally refer to cultural values, and (vii) - (x) to natural ones. At least one of the ten criteria must be fulfilled by any nomination to prove its "*exceptional [significance] as to transcend national boundaries and to be of common importance for present and future generations of all humanity*" (UNESCO, 1972, Jokilehto, 2008). Since 2007, a complete Statement of OUV is required for new nominations to contain brief synthesis, justification for criteria, statement of integrity and/or authenticity, and requirements for protection and management. The section **justification for criteria** explains why a property fulfills all criteria under which it has been inscribed, giving a concise paragraph for each criterion. Retrospective Statements of OUV were also prepared during the Second Cycle of Periodic Reporting (2008-2015) by 812 properties[1] inscribed before 2006, to revise or refill the section of justification for criteria if it was incomplete or not agreed on at the time of inscription (IUCN et al., 2010).

[1] this number is calculated based on the data provided in the Reports of each regions available at http://whc.unesco.org/en/pr-questionnaire/

Investigating OUV and comparing it to the selection criteria and justifications applied to the listed WH properties is not uncommon. Most research, however, focuses on a single case or a few cases for comparative study, thus mainly concerning a small number of Statements of OUV (Shah, 2015, Ruffino et al., 2019, Abdel Tawab, 2019, Tarrafa Silva and Pereira Roders, 2010). Whereas the 2007 *International Conference on Values and Criteria in Heritage Conservation* explicitly organized sessions to discover the definition and evolution of OUV as an integral concept, discussing the terms used in the current (by then) WH justifications and proposing possible enhancement to clarify the concepts (Fejérdy, 2007, Petzet, 2007, Jokilehto, 2007). The whole discussion of this conference resulted in the well-known ICOMOS report "*What is OUV, Defining the Outstanding Universal Value of Cultural World Heritage Properties*", published in 2008. The report described the evolution of OUV since first proposed, summarized the essential focuses of each cultural selection criterion, and matched the criteria to the main themes in existing WH properties (Jokilehto, 2008). In that report, the concepts of OUV are illustrated from both a deductive perspective by interpreting the definitions in *Operational Guidelines*, and an inductive perspective by giving examples from justification texts of WH properties. Keywords in the justifications are highlighted to indicate why this piece of text reflects the selection criterion it describes. Furthermore, the report suggests that the criteria are strongly associated with each other, since that the "*historical value is an integral part of the majority of... criteria (i)-(vii)*", and that "*the aesthetic /artistic value also plays a role in several OUV criteria*". Such associations have been further investigated in the report by looking at how often a specific criterion is used together with the others.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-VIII-M-1-2021-25-2021 | © Author(s) 2021. CC BY 4.0 License.

This line of interpreting OUV and the selection criteria is rather effective and contributes to a better understanding of the concepts. However, such processes of keywords highlighting are heavily dependent on expert knowledge, which may not be easily applicable and intelligible for the general public, let alone being prone to inevitable personal and disciplinary biases. A recent study took all the available Statements of OUV in the World Heritage List (concerning 1049 properties that have a complete section of justification for criteria) as input data and trained several state-of-the-art Natural Language Processing (NLP) models on an OUV classification task (Bai et al., 2021). That study revealed a top-3 accuracy of 94% to predict the correct selection criterion, based on the short piece of text justifying this criterion. The authors also provided an open-source repository with all their trained models and results concerning the models' performances[2]. This previous study provides a chance to revisit the 2008 ICOMOS report from a computational perspective to re-interpret the focuses, definitions, and associations of the selection criteria that define the OUV of WH properties.

This paper presents a computational analysis of the selection criteria justified in Statements of OUV on their semantic meanings and intrinsic associations. The contributions can be summarized as: 1) providing an OUV-related lexicon that can be used to highlight keywords in a generic text on relevant selection criteria; 2) proposing three types of matrix-based similarity metrics from different sources to represent the pair-wise associations of criteria; 3) conducting qualitative and quantitative analyses on the lexicon and the similarity metrics, which may give insights to more clearly defining OUV in future practice.

2. METHODOLOGY

2.1 Input Materials and Problem Statement

The following variables $\mathbf{A}, M, \mathbf{C}^{(i,s)}$, and $W_k^{(i)}$ are derived from the open-source repository of the study mentioned in Section 1 and are applied as the input material for this study.

Considering all the properties inscribed in the WH List, a co-occurrence matrix of the selection criteria was constructed as $\mathbf{A} = [A_{k,l}]_{\kappa\times\kappa},\ k,l \in [0,\kappa), \kappa = 10$, where the off-diagonal entries $A_{k,l}, k \neq l$ are the number of properties that satisfy both criteria k and l, and the diagonal entries $A_{k,k}$ record the number of cases when each criterion k is used alone (see Figure 1a).

Five state-of-the-art NLP models $M = \{m_i | i = [0,5)\}$ were trained and tested on classifying selection criteria from sentences, which stand for N-Gram (Cavnar and Trenkle, 1994), Bag-of-Embeddings (Pennington et al., 2014), Attention with GRU (Yang et al., 2016), BERT (Devlin et al., 2019), and ULMFiT (Howard and Ruder, 2018), respectively. The latter two were proved to perform better in terms of classification accuracy. For each model m_i, three confusion matrices $\mathbf{C}^{(i,s)} = [C_{k,l}^{(i,s)}]_{\kappa\times\kappa},\ k,l \in [0,\kappa), s \in \{\text{train, val, test}\}$ were provided, where the entries $C_{k,l}^{(i,s)}$ represent the total number of data samples with a true label of criterion k being classified as criterion l by model m_i in the s set (train, validation, or test datasets). An example of the confusion matrix $\mathbf{C}^{(4,\text{test})}$ of m_4's (ULMFiT) performance on test dataset is shown in Figure 1b.

A total of 2353 phrases composed of 1- to 5-Gram features (phrases with 1 to 5 consequent words) that appeared more than 15 times and less than 600 times in the Statements of OUV were fed to each model, predicting the scores of each phrase belonging to each criterion $k, k \in [0, \kappa+1)$, where the 11_{th} criterion referred to an additional negative class of "Others" related to none of the criteria. A series of ordered sets $W_k^{(i)} = \{(\text{phrase } w, \text{rank } r)\}, |W_k^{(i)}| = 50, r \in [1,50]$ of phrases was obtained to contain the ranked top-50 keywords for criterion k predicted by the model m_i. The initial vocabulary can be composed of all the phrases as $V^{(0)} = \bigcup_{k=0}^{\kappa+1}\bigcup_{i=0}^{5}\{w|(w,*) \in W_k^{(i)}\}, |V^{(0)}| = 1782$. A three-dimensional array $\boldsymbol{\Upsilon} = [\upsilon_{j,k,i}]_{|V^{(0)}|\times(\kappa+1)\times 5}$ can be constructed for the j_{th} phrase w_j in the vocabulary $V^{(0)}$ pertaining to its rank r in the criterion k predicted by model m_i, such that:

$$\upsilon_{j,k,i} = \begin{cases} r, & \text{if } (w_j, r) \in W_k^{(i)}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The above-mentioned variables and the processed $V^{(0)}$and $\boldsymbol{\Upsilon}$ are used to construct the lexicon and the similarity graphs.

2.2 Keywords Lexicon

Lexicon, literally defined as "*all the words and phrases used in a particular language or subject*"[3] was originally a linguistic concept, which requires some "*morpholexical rules*" to specify whether words should be members of some classes (Lieber, 1980). However, in modern NLP literature, the term "lexicon" is frequently referred to as a list of words that "*carry particularly strong cues*" of certain word senses, usually sentiment (Jurafsky and Martin, 2020, Faruqui et al., 2015). One of the most popularly used lexicons is the SentiWordNet, where each word is given scores for its tendency of being positive, negative, and objective (Esuli and Sebastiani, 2006). Such lexicons can be constructed by manual annotation, semi-supervised induction, and/or supervised learning. The initial entire vocabulary $V^{(0)}$ has the following problems to be considered as a lexicon, which needs to be revised and filtered: 1) some terms only appear in a limited number of models (especially in the worse performing models such as m_1 N-Gram model), which may be caused by the randomness of the models (e.g., "*foot*" was predicted with a high rank by m_1); 2) some terms always have lower confidence scores (lower ranks) in all models, which may suggest that they are not strongly relevant to the topic; 3) some terms are redundant since the longer N-Gram features may be accompanied by their subsets, for example "*directly and tangibly associated*" appears together with "*directly and tangibly*", "*and tangibly associated*", etc.; 4) stop-words such as prepositions and articles differentiate the word senses in their contexts (Devlin et al., 2019), but may not introduce additional semantic meanings when considered as keywords (e.g., "*art of*", "*art in*", and "*art and*" are all about the concept "*art*").

To improve these aspects, keywords are aggregated by taking advantage of the ensemble of models. Since the performance of the model may suggest the general reliability of predicted keywords, a model-related weight vector $\boldsymbol{\omega} = [\omega_i]_{5\times 1} = [1,1,1,\lambda_0,\lambda_0]^T, \lambda_0 \geq 1 \in \mathbb{R}^+$ is arbitrarily formed to give the predictions by the latter two models a higher weight. Similarly, keywords predicted with higher confidence scores (higher ranks) may suggest that they are more related to the topic. Therefore, a rank-related weight vector $\boldsymbol{\zeta} = [\zeta_r]_{51\times 1} =$

[2] https://github.com/zzbn12345/WHOSe_Heritage, Copyright (c) 2021 Nan BAI & Renqian LUO under The MIT License

[3] Oxford Learner's Dictionary

$[0, \lambda_1^2, ..., \lambda_1^2, \lambda_1, ..., \lambda_1, 1, ..., 1]^T, \lambda_1 \geq 1 \in \mathbb{R}^+$ is also arbitrarily constructed to give higher-ranked keywords more importance, where the top-10 are amplified by the scalar λ_1^2, the $11_{\text{th}} - 25_{\text{th}}$ ranked phrases are amplified by λ_1, the $26_{\text{th}} - 50_{\text{th}}$ are kept the same, and those not ranked are omitted. The three-dimensional array $\boldsymbol{\Upsilon}$ in equation 1 can be therefore flattened on the model axis i to a matrix $\boldsymbol{\Upsilon'} = [v'_{j,k}]_{|V_0| \times (\kappa+1)}$, such that:

$$v'_{j,k} = \sum_{i=0}^{4} \zeta[v_{j,k,i}]\omega[i]. \tag{2}$$

With a threshold $\lambda_2 \in \mathbb{R}^+$ to filter the computed weights in the matrix $\boldsymbol{\Upsilon'}$, a group of aggregated keyword sets W'_k can be obtained for each criterion k, such that:

$$W'_k = \{(w_j, v'_{j,k}) | v'_{j,k} \geq \lambda_2\}. \tag{3}$$

Finding a properly filtered group of sets W'_k can be formulated as the following optimization problem, where W'_k is effectively a function of the three variables $\lambda_0, \lambda_1, \lambda_2$:

$$\max_{\lambda_0, \lambda_1, \lambda_2} \quad \frac{|\bigcup_{\substack{k,l=0 \\ k \neq l}}^{\kappa+1} (\{w|(w,*) \in W'_k\} \cap \{w|(w,*) \in W'_l\})|}{|\bigcup_{k=0}^{\kappa+1} \{w|(w,*) \in W'_k\}| \times \sigma_{|W'_k|} + \epsilon}, \tag{4a}$$

$$\text{subject to} \quad |\bigcup_{k=0}^{\kappa+1} \{w|(w,*) \in W'_k\}| \leq N_0 = 800 \tag{4b}$$

$$\lambda_0, \lambda_1, \lambda_2 \in \{1.0, 1.1, 1.2, \ldots, 4.9\}. \tag{4c}$$

Where $\sigma_{|W'_k|}$ denotes the standard deviation of the sizes of sets W'_k, and ϵ is a small number to avoid zero division. This optimization ensures that: 1) there are enough phrases that fulfill more than one criteria (ensured by the nominator of equation 4a); 2) the total size of the vocabulary is concise (ensured by N_0 in equation 4b); 3) the sizes of keyword sets are evenly distributed across the criteria (ensured by $\sigma_{|W'_k|}$ in the denominator of equation 4a); and 4) the weights are in reasonable ranges for the filtering computation (ensured by equation 4c).

Using a brute-force search for solving this optimization from a total of $|\lambda_0||\lambda_1||\lambda_2| = 64000$ configuration possibilities of discretized $\lambda_0, \lambda_1, \lambda_2$, a configuration of $\lambda_0 = 2.2, \lambda_1 = 1.2, \lambda_2 = 2.6$ yields the best filtering with a total vocabulary size of $|V^{(1)}| = |\bigcup_{k=0}^{\kappa+1}\{w|(w,*) \in W'_k\}| = 552$, among which 78 occur in more than one selection criteria. For the new vocabulary $V^{(1)}$, Stop-words and WordNet Lemmatizer tools in the NLTK package (Loper and Bird, 2002, Miller, 1995) are used to further normalize and merge the keywords (as with the example of "*art*"). Furthermore, phrases composed of more than 2 words are merged to their longest N-Gram features (as with the example of "*directly and tangibly associated*"). After merging, a final lexicon as sets W_k is obtained, yielding a vocabulary size of $|V| = |\bigcup_{k=0}^{\kappa+1}\{w|(w,*) \in W_k\}| = 354$, among which 77 occur in more than one selection criteria.

2.3 Similarity Matrices

Co-occurrence matrix $\mathbf{A}$ of the selection criteria, as introduced in Section 2.1, shows how often two criteria are justified together, i.e. marked as relevant, for a WH property. The more often two criteria are fulfilled simultaneously, the more similar and associated they arguably are with one another. The term "similarity" here is from a **structural** viewpoint on the dataset. By normalizing matrix $\mathbf{A}$, the upper triangular entries can be "unrolled" and form a long vector $\boldsymbol{\alpha} = [\alpha_t]_{\frac{\kappa(\kappa-1)}{2} \times 1}, t \in [0, \frac{\kappa(\kappa-1)}{2})$, indexed with the ordered pair $(k, l), k < l$, representing the pair-wise similarity of the criteria, such that:

$$\{\alpha_t\} = \left\{ \frac{\kappa A_{k,l}}{\sum_{k_0} \sum_{l_0} A_{k_0,l_0}} | k, l \in [0, \kappa), k < l \right\}. \tag{5}$$

On the other hand, the confusion matrices $\mathbf{C}^{(i,s)}$ of the models during training and testing processes reveal how easily different selection criteria are to be misclassified as each other. Suppose the models are properly trained and represent certain degrees of truth, two criteria shall be more similar to one another as the models literally "confuse" them more often (Zhang et al., 2019). The term "similarity" here is an **experimental** viewpoint on the data concerning the NLP models' performances. However, before arguing that the confusion matrices reflect some intrinsic similarity, one must first prove that the models behave in a consistent manner, i.e., different models have difficulties at the same criteria pairs by easily confusing them. For each combination of the performance of model m_i on either validation or test set s (training set performances are disregarded since the other two are supposed to better represent the prediction power of models), a similar construction as equation 5 can be applied to obtain long vectors $\boldsymbol{\beta}^{(i,s)} = [\beta_t^{(i,s)}]_{\frac{\kappa(\kappa-1)}{2} \times 1}, t \in [0, \frac{\kappa(\kappa-1)}{2})$ from the confusion matrices $\mathbf{C}^{(i,s)}$ following (Zhang et al., 2019), such that:

$$\{\beta_t^{(i,s)}\} = \left\{ \frac{C_{k,l}^{(i,s)}}{\sum_{k_0} C_{k_0,l}^{(i,s)}} + \frac{C_{l,k}^{(i,s)}}{\sum_{l_0} C_{l_0,k}^{(i,s)}} | k, l \in [0, \kappa), k < l \right\}. \tag{6}$$

Since the co-occurrence matrix $\mathbf{A}$ is symmetrical, the summation in Equation 6 is desirable as it transforms the generally asymmetrical confusion matrices into symmetric ones. The long vectors $\boldsymbol{\beta}^{(i,s)}$ are first compared to each other using Spearman's Rank Correlation to check the consistency of the models' performances. However, the null hypotheses in normal correlation analyses on such vectors can be easily refuted falsely because of the auto-correlated structures in matrices, making the normal significance tests invalid. A method called Quadratic Assignment Procedure (QAP) has been proposed to solve this problem (Liu, 2007, Krackhardt, 1988). By repeating the process of simultaneously permuting the rows and columns of one of the matrices before unrolling it to a vector for correlation computation, a theoretical distribution of the correlation coefficients can be obtained as a simulation outcome. The percentile of the original correlation coefficient (the one calculated without permutation) in this theoretical distribution can instead estimate the significance level of the correlation analyses effectively. The vectors are then fed to Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF) algorithms in Scikit-learn to perform dimensionality reduction and obtain the aggregated vector $\boldsymbol{\beta} = [\beta_t]_{\frac{\kappa(\kappa-1)}{2} \times 1}, t \in [0, \frac{\kappa(\kappa-1)}{2})$, representing the pair-wise confusion of the selection criteria (Févotte and Idier, 2011).

Furthermore, the final lexicon $V = \bigcup_{k=0}^{\kappa+1}\{w|(w,*) \in W_k\}$ discussed in section 2.2 can provide another level of interpretation on the criteria similarity. As suggested by the NLP literature (Pennington et al., 2014, Mikolov et al., 2013, Wal-

lach, 2006), the pre-computed word embedding vectors provide good semantic meanings of the phrases, which can be further aggregated to represent the document topics composed of the ensemble of words. Therefore, another matrix $\mathbf{H} = [H_{k,l}]_{\kappa\times\kappa},\ k,l \in [0,\kappa)$ showing the **semantic** similarity of the criteria can be constructed by computing the pair-wise cosine similarities of the averaged embedding vectors $\mathbf{f}_k$ of phrases in W_k for each criterion k, such that:

$$\mathbf{f}_k = \frac{\sum_{j=0}^{|V|}\mathbf{g}(w_j)}{|W_k|}|(w_j, v'_{j,k}) \in W_k. \tag{7}$$

Where $\mathbf{g}(w_j)$ is a function to look up the 300-dimensional GloVe embedding vectors of all the words in the phrase w_j and take the sum of the vectors. Similar to equation 5, another long vector $\boldsymbol{\gamma} = [\gamma_t]_{\frac{\kappa(\kappa-1)}{2}\times 1}, t \in [0, \frac{\kappa(\kappa-1)}{2})$ can be obtained to represent the pair-wise semantic similarities of the criteria.

$$\{\gamma_t\} = \left\{ H_{k,l} = \frac{\mathbf{f}_k^T\mathbf{f}_l}{||\mathbf{f}_k||_2||\mathbf{f}_l||_2}|k,l \in [0,\kappa), k<l \right\}. \tag{8}$$

The three vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are further compared to each other using Spearman's Rank Correlation (as they have different value distributions) to check the relationship and consistency of different similarity definitions based on QAP significance level.

2.4 Graph Visualization

The vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ representing the pair-wise similarity of the selection criteria can be also interpreted as the edge weights of three undirected weighted unipartite graphs $G_\alpha, G_\beta, G_\gamma$, where each node represents a specific criterion k. The graphs are visualized in Gephi using the Force Atlas algorithm based on the edge weights (Bastian et al., 2009, Jacomy et al., 2014). Since those graphs are (almost) complete with significantly divergent edge weights, different thresholds $\xi_\alpha, \xi_\beta, \xi_\gamma$ are applied to show only the edges whose weights are larger than the threshold based on the weight distributions, in order to give clearer structural information of the associations between the criteria.

Furthermore, the lexicon, i.e., the ensemble of sets $\bigcup_{k=0}^{\kappa+1} W_k = \{(w_j, v'_{j,k})\}$ can also be interpreted as the edge table of an undirected weighted bipartite graph B_w, where the two sets of nodes are respectively the vocabulary V and all the selection criteria. Moreover, as introduced in section 2.2, some phrases may belong to more than one criteria, and edge weights of such phrases can also vary across criteria. For example, the term "*architectural*" belongs to both Criterion (iv) with a weight of 5.70 and Criterion (i) with a weight of 4.75. In such cases, the degree of nodes representing the phrases will be the sum of weights from all edges connected to them. The lexicon as a bipartite graph is also visualized in Gephi using the Force Atlas algorithm based on the edge weights.

3. RESULTS

3.1 OUV-related Lexicon of Selection Criteria

The visualized lexicon as bipartite graph B_w containing all phrases in V and their relationship with the selection criteria (including the negative class "Others") are shown in Figure 2. Generally, the essential topics of the criteria also appear to have the largest weights as the prediction from computational models. This is obvious in the cases of Criterion (i) with phrase

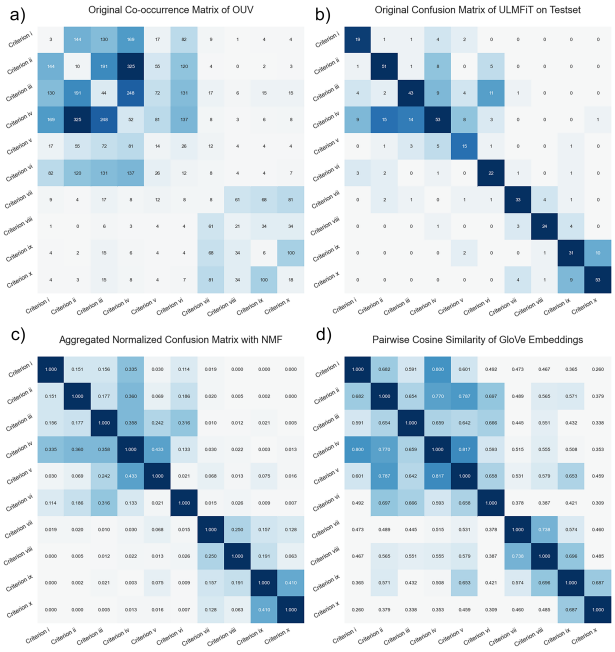

Figure 1. The matrices representing the pairwise similarity and associations between selection criteria. a) the original (unnormalized) co-occurrence matrix $\mathbf{A}$; b) the original (unnormalized) confusion matrix $\mathbf{C}^{(4,\text{test})}$ by m_4 ULMFiT; c) the aggregated normalized confusion constructed from the NMF vector $\boldsymbol{\beta}$; d) the semantic similarity matrix $\mathbf{H}$ of the pairwise cosine similarity of GloVe embeddings for each criterion.

"*masterpiece*" and "*human creative genius*", (ii) with "*influence*" and "*development*", (iii) with "*bear exceptional testimony*", (iv) with "*outstanding example*" and "*building*", (v) with "*traditional human settlement*", (vi) with "*directly and tangibly associated*", (vii) with "*exceptional natural beauty*", (viii) with "*geological process*", (ix) with "*ecological*", and (x) with "*species*". For each criterion, not only adjectives and verb phrases describing the **values**, but also nouns and noun phrases showing the critical **attributes** can be found. Take Criterion (i) as an example, phrases such as "*unique artistic achievement, creative, genius, artistic, monumental*" highlight the main artistic, aesthetic, and historic values associated with this criterion. Meanwhile representative attributes such as "*fresco, sculpture, interior, decoration, art and architecture*" demonstrate where those values are applied to.

Inspecting the phrases associated with more criteria can provide some insights into interpreting the common justifications of OUV. The terms "*art*" and "*design*" connect Criteria (i)(ii)(iv), while "*landscape*" connects Criteria (i)(ii)(v), and "*cultural landscape*" connects Criteria (iv)(v), showing the common stand-points and nuances in the focuses of those criteria. Moreover, the groups of phrases related to religions connecting Criteria (iii) and (vi), phrases about architectural art connecting (i) and (iv), about urban form connecting (iv) and (v), about natural phenomena between (vii) and (viii), as well as phrases about bio-creatures between Criteria (ix) and (x), etc., all imply some common characteristics within the OUV concept.

3.2 Matrix Similarities

All vector pairs from $\boldsymbol{\beta}^{(i,s)}$ have a high Spearman's Rank Correlation coefficient from 0.713 to 0.933, while all correlations are significant with $p < 0.001$ based on QAP simulation. This

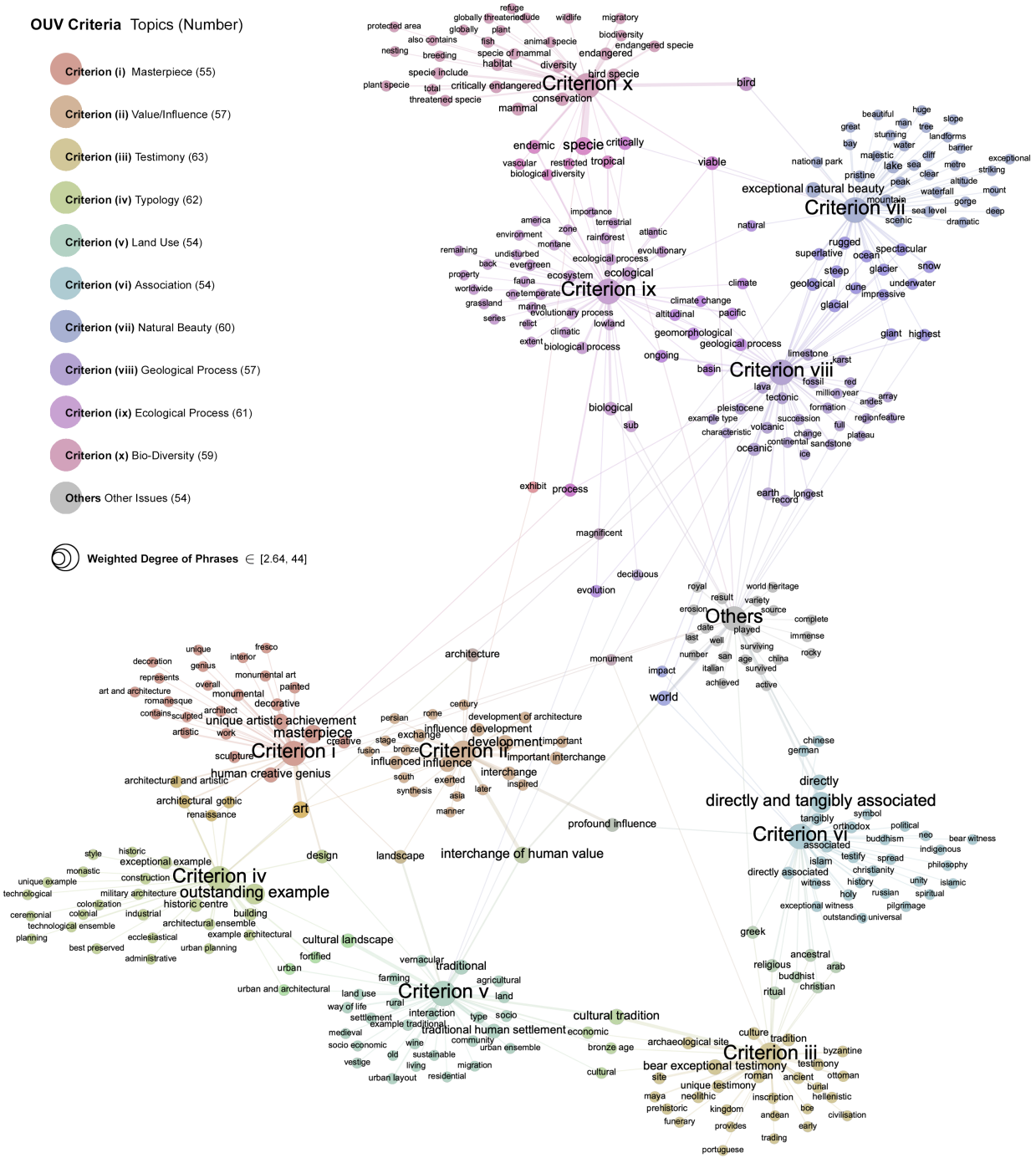

Figure 2. The lexicon of selection criteria, i.e., the bipartite graph B_w, visualized as a word network based on the Force Atlas algorithm in Gephi. Thicker edges indicate higher weights of the phrases in vocabulary V regarding a specific criterion. Nodes with higher edge weights are placed closer to each other in the visualization. Larger nodes and font sizes indicate larger total weighted degrees of the phrases. The colors of the phrase nodes are rendered the same as the criterion they belong to. The nodes of phrases belonging to two or more criteria are placed between the criteria clusters, and the colors of the nodes are also the mixture of the criteria colors. The general topics of criteria according to the ICOMOS report (Jokilehto, 2008) and the total number of keywords belonging to each criterion, i.e., $|W_k|$ are demonstrated in the legend. This graph (lexicon) could be used to locate specific words regarding their relations with different selection criteria, and to observe and select the most relevant words while drafting and/or evaluating the Statements of OUV. Detailed interpretations of the lexicon are presented in Section 3.1.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-VIII-M-1-2021-25-2021 | © Author(s) 2021. CC BY 4.0 License.

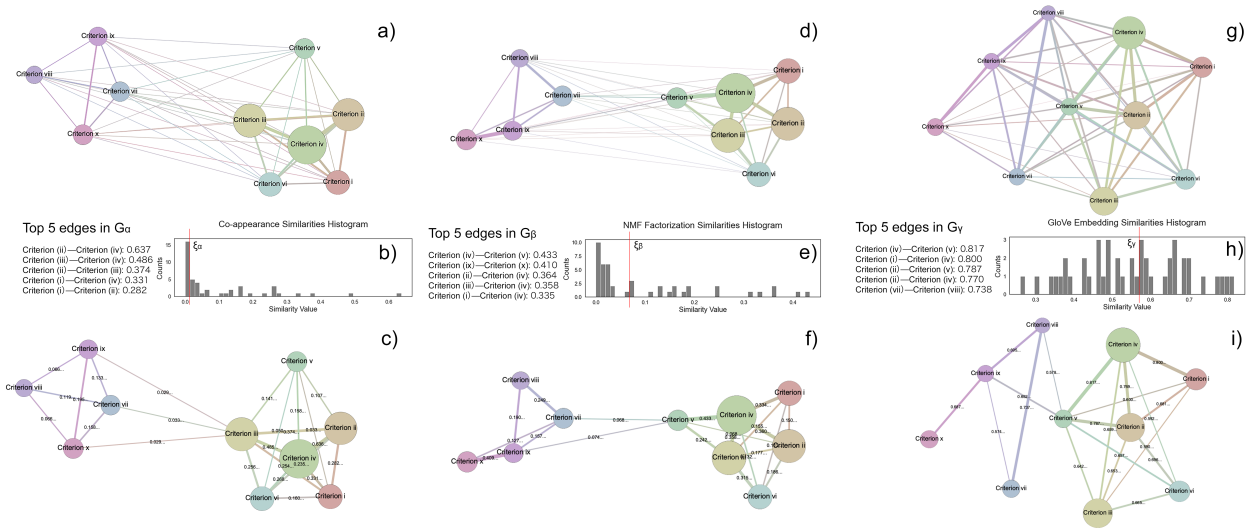

Figure 3. The graph visualizations of the similarity matrices represented by $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ as edge weights using the Force Atlas algorithm in Gephi. a-c) Co-occurrence graph G_α; d-f) Confusion graph G_β; g-i) Semantic similarity G_γ; a/d/g) Complete graphs with all edge weights visualized; c/f/i) Filtered graphs that only show edges whose weights are higher than the first two cross-domain cultural-natural criteria pair; b/e/h) Histogram of edge weights and the threshold $\xi_\alpha, \xi_\beta, \xi_\gamma$ during filtering, the top-5 edges being listed with their weights. Node size represents the total World Heritage properties justified with this selection criterion.

Vector 1	**Vector 2**	ρ **value**	p **value**
$\boldsymbol{\alpha}$ (Structural)	$\boldsymbol{\beta}$ (Experimental)	0.838*	$<.001$
$\boldsymbol{\alpha}$ (Structural)	$\boldsymbol{\gamma}$ (Semantic)	0.615*	$<.001$
$\boldsymbol{\beta}$ (Experimental)	$\boldsymbol{\gamma}$ (Semantic)	0.793*	$<.001$

*$p < .001$ with QAP simulation of 1000 permutations.

Table 1. The Spearman’s Rank Correlation of three long vectors. The significance level p is computed based on QAP simulation.

suggests that all the investigated confusion matrices perform consistently across models and datasets. Though models such as BERT and ULMFiT generally have a better prediction accuracy, they are similarly confused at the same criteria pairs as the worse-performing models. Therefore, it is appropriate to aggregate the vectors $\boldsymbol{\beta}^{(i,s)}$ into $\boldsymbol{\beta}$ to represent the overall confusion patterns of the models. The first PCA component of the vectors manages to explain 89.7% of the variance in $\boldsymbol{\beta}^{(i,s)}$. However, due to the nature of PCA, some elements in its component are unavoidably negative, which can be hard to interpret as a similarity metric. Alternatively, the first component computed from NMF is non-negative, and has a Pearson Correlation of $r = 1.0$, $p < 0.001$ with the first PCA component. Therefore, the first NMF component from $\boldsymbol{\beta}^{(i,s)}$ is used as $\boldsymbol{\beta}$ for later analysis. This vector effectively makes a single matrix representative of the 10 possible variants of the G_β, thus making this graph comparable to the other two graphs.

The values of the vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are reflected in Figure 1 (a), (c), and (d), respectively. The matrix heatmaps generally illustrate a consistent visual pattern: 1) the top left corner indicating the cultural criteria associations and the bottom right corner indicating the natural criteria associations are stronger and create two relatively dense sub-matrices; 2) the off-diagonal entries highlight similar places, such as the entries representing the relation between Criteria (ii)(iv) and between Criteria (ix)(x). These patterns are further proved with correlation analysis. The Spearman’s Rank Correlation of the vectors representing the similarities between selection criteria is shown in Table 1. All three pairs are significantly correlated with a high coefficient between 0.615 and 0.838, proving that the three proposed similarity matrices representing the structural (as co-occurrence matrix), experimental (as aggregated confusion matrix), and semantic (as cosine similarity matrix of GloVe embedding) information of the criteria are consistent with each other, though each one of the three may capture different aspects of the pairwise associations. These aspects will be discussed extensively in Sections 3.3 and 4. The QAP-simulation-based p values out of 1000 random permutations indicate that such high correlations are significant, i.e. not caused by randomness.

3.3 Associations and Similarities of Selection Criteria

The similarity matrices showing the associations of selection criteria are further visualized in 2D as weighted graphs $G_\alpha, G_\beta, G_\gamma$ in Figure 3, where the nodes representing more similar criteria are placed closer to each other. The graphs on the top are complete graphs showing all edge weights, while the graphs on the bottom are filtered graphs only showing the edges whose weights are equal or higher than the first two cross-domain edges linking cultural (i-vi) and natural (vii-x) criteria. The thresholds $\xi_\alpha, \xi_\beta, \xi_\gamma$ for conducting the filtering are also plotted on the histograms of the edge weights. It can be observed from the histograms that the edge weights in G_α and G_β are more divergent, while in G_γ, the edge weights are more homogeneous. As a consequence, G_γ is also visually more different from the other two similarity graphs.

By inspecting the visualization in Figure 3, consistent association and similarity patterns of the criteria can be observed from the graphs: 1) the in-domain edges generally have a larger weight than cross-domain edges, thus creating two sub-graph clusters for cultural and natural criteria in all graphs, suggesting that cultural and natural criteria are relatively independent with each other; 2) the first several cross-domain edges connecting cultural and natural criteria always involve either Criterion (v) about *Land-Use* or Criterion (iii) about *Testimony*, suggesting that these two cultural criteria also have a natural aspect; 3) the cultural criteria are generally more connected and interrelated than the natural ones, suggesting that the cultural cri-

teria are probably more similarly defined and associated with each other than the natural criteria; 4) the edges between Criteria (ii) and (iv), and between Criteria (i) and (iv) are always among the top-5 weights in all three graphs (see the lists of Top 5 edges in Figure 3b/e/h), proving the strong association of *Architectural Typology* with both *Masterpiece* and cultural *Influences*; 5) the edge between Criteria (iv) and (v) appears to be the top-1 weight of both G_β and G_γ, but is only the 13_{th} in G_α, showing that the association of *Architectural* heritage and *Urban* heritage might be stronger than indicated by the actual co-justification in WH list; 6) Contrarily, the edges between Criteria (iii) and (iv), and between Criteria (ii) and (iii) are ranked top-3 in graph G_α, yet respectively rank as 11_{th} in G_γ and G_β, showing that although these criteria are usually co-justified in WH properties, they may not be that semantically similar or empirically confusing.

Remarkably, the strong associations indicated by the graphs in Figure 3 are also clearly illustrated with many common phrases (lexicon) in Figure 2, though the two figures are derived from different data sources and resolutions. The bipartite lexicon graph B_w in Figure 2 can be interpreted more as a zoomed-in view on the selection criteria composed of phrases, while the graphs $G_\alpha, G_\beta, G_\gamma$ in Figure 3 arguably reflect a zoomed-out view on the characteristics of criteria themselves.

4. DISCUSSION

The lexicon presented in Figure 2 could become a tool for researchers and practitioners to automatically highlight the keywords in a sentence about World Heritage properties and indicate the best matching selection criteria, which also has the potential to facilitate the drafting and revising of SOUV, useful to support new WH nominations and their evaluation by the Advisory Body Evaluation parties, ICOMOS and IUCN. Since the computational models were trained with the authoritarian context of WH properties, the lexicon derived from this study provides a chance to empirically investigate the patterns frequently appeared in Statements of OUV which are captured and learned by the NLP models, while they can be easily neglected or undervalued with traditional methods. For example, Criterion (i) is officially defined as "*to represent a masterpiece of human creative genius*" in the *Operational Guidelines* and summarized as "*masterpiece*" by the 2008 report (UNESCO, 2008, Jokilehto, 2008). However, the term "*unique artistic achievement*" is boldly stressed by the computational models and the lexicon shown in Figure 2, suggesting that artistic value is also expected to be of high importance for the WH properties justified with Criterion (i). Similarly, though Jokilehto stressed more on the "*value/influence*" dimension of Criterion (ii), the terms related to "*development*" and "*interchange*" in its definition also seem to have alike importance. As the next step, the lexicon could be further updated with additional human engineering such as expert-based rating, as the current version is the outcome of a semi-automated procedure. Although filtering as described in Section 2.2 has been applied, not every phrase in the lexicon makes sense. Some failure examples include the term "*one*" and "*back*" within Criterion (ix), "*total*" within Criterion (x), and "*overall*" within Criterion (i). Those terms should have been rather neutral, but probably the consistent writing style and word usage preference in Statements of OUV give some phrases a misleading score. Furthermore, the lexicon can be used as initial "seed words" in future studies to construct a more comprehensive and concrete World Heritage OUV-related lexicon by incorporating other larger and maturer semantic lexicons such as WordNet (Miller, 1995, Jurafsky and Martin, 2020).

Some visual similarities can already be observed in Figure 1, as the heatmaps seem to highlight matrix entries in a similar pattern. This was also probably the assumption in ICOMOS 2008 report about the OUV associations, as argued in Section 1. Yet these similarities would be hard to prove and falsify without a quantitative methodology, such as the one presented in this paper. The correlation coefficients shown in Section 3.2 and the graphs $G_\alpha, G_\beta, G_\gamma$ in Figure 3 confirm this intuitive assumption based on observations. Furthermore, while graph G_α based on the co-occurrence pattern of the OUV criteria may vary radically due to the change of interest or focus of the WH Committee during the nomination procedure, the other two graphs might be more static along the time. The 2008 ICOMOS report argued that "[Criteria] *(i) and (ii) can reinforce each other, while (iv) is often used as an alternative*" based on the co-occurrence pattern at that time, when cases co-justifying Criteria (i) and (ii) were almost twice as many as the cases with Criteria (i) and (iv) (Jokilehto, 2008). This observation is no longer true for the situation in 2019, when the latter, i.e. cases with Criteria (i) and (iv), appears even more frequently than the former. However, both associations are observed in the 4_{th} finding presented in Section 3.3. As graph G_β and G_γ are both based on the written texts and terms collectively used in the entire Statements of OUV, they may be more robust to new nominations unless very unusual terms are to be systematically introduced. It can also be informative in future studies to investigate the changing dynamic of presented graphs along time.

The qualitative and quantitative analyses show that the selection criteria pairs have different association strengths. For a thoroughly trained expert (either human or computer), nuances between pairs such as Criteria (i) and (iv) can already be rather hard to distinguish, let alone someone from the general public. To make the World Heritage management more socially inclusive, the concept of OUV more intelligible, and the future inscription process more effective, extra efforts may need to be made to further sharpen and clarify the definitions of criteria, and to make sure the OUV statements written by future practitioners and researchers are sufficiently consistent and coherent.

5. CONCLUSIONS

This paper presents the computational interpretation on the associations of UNESCO World Heritage selection criteria indicating the Outstanding Universal Value (OUV) conveyed by the properties, as an evolution of the ICOMOS report "*What is OUV*" published in 2008, applying a novel methodology integrating state-of-the-art technology. It provides an OUV-related lexicon showing relevant phrases of each selection criterion, proposes three similarity graphs using different data sources to show various aspects of the criteria associations, and conducts quantitative and qualitative analyses on the lexicon and similarity graphs to make sense of the observations. This study may give some insights to further evolutions and improvements of the concept of both World Heritage and OUV, as is also regularly revised by the World Heritage Committee[4].

ACKNOWLEDGEMENTS

The presented study is within the framework of the Heriland-Consortium. HERILAND is funded by the European Union's

[4] http://whc.unesco.org/en/criteria/

Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 813883.

REFERENCES

Abdel Tawab, 2019. The Assessment of Historic Towns' Outstanding Universal Value Based on the Interchange of Human Values They Exhibit. *Heritage*, 2(3), 1874–1891.

Bai, N., Luo, R., Nourian, P., Roders, A. P., 2021. WHOSe Heritage: Classification of UNESCO World Heritage "Outstanding Universal Value" Documents with Smoothed Labels. *arXiv preprint arXiv:2104.05547*.

Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: an open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 3, 361–362.

Cavnar, W. B., Trenkle, J. M., 1994. N-gram-based text categorization. *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, 161–175.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Esuli, A., Sebastiani, F., 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. *LREC*, 6, 417–422.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., Smith, N. A., 2015. Retrofitting word vectors to semantic lexicons. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1606–1615.

Fejérdy, T., 2007. Evolution and possible enhancement of the concept of ouv. *Values and Criteria in Heritage Conservation*, Polistampa, 323–328.

Févotte, C., Idier, J., 2011. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9), 2421–2456.

Howard, J., Ruder, S., 2018. Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339.

IUCN, ICOMOS, ICROM, WHC, 2010. Guidance on the preparation of retrospective Statements of Outstanding Universal Value for World Heritage Properties. Technical report.

Jacomy, M., Venturini, T., Heymann, S., Bastian, M., 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6), e98679.

Jokilehto, J., 2007. Aesthetics in the world heritage context. *Values and Criteria in Heritage Conservation*, Polistampa, 183–194.

Jokilehto, J., 2008. What is OUV? Defining the Outstanding Universal Value of Cultural World Heritage Properties. Technical report, ICOMOS, ICOMOS Berlin.

Jurafsky, D., Martin, J. H., 2020. Speech and language processing An Intoduction to Natural Language Processing, Computational Liguistics, and Speech Recognition (Third Edition Draft).

Krackhardt, D., 1988. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social networks*, 10(4), 359–381.

Lieber, R., 1980. On the organization of the lexicon. PhD thesis, Massachusetts Institute of Technology.

Liu, J., 2007. QAP: A Unique Method of Measuring "Relationships" in Relational Data. *Chinese Journal of Sociology(in Chinese Version)*, 27(4), 164-174.

Loper, E., Bird, S., 2002. Nltk: the natural language toolkit. *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, 63–70.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. *ICLR*.

Miller, G. A., 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39–41.

Pennington, J., Socher, R., Manning, C. D., 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Petzet, M., 2007. What is outstanding universal value. *Values and Criteria in Heritage Conservation*, Polistampa, 315–322.

Ruffino, P., Permadi, D., Gandino, E., Haron, A., Osello, A., Wong, C., 2019. Digital technologies for inclusive cultural heritage: The case study of serralunga d'alba castle. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 141–147.

Shah, K., 2015. Documentation and cultural heritage inventories case of the historic city of ahmadabad. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2, 271–278.

Tarrafa Silva, A., Pereira Roders, A., 2010. The cultural significance of World Heritage cities : Portugal as case study. *Heritage and Sustainable Development*, Évora, Portugal.

UNESCO, 1972. Convention Concerning the Protection of the World Cultural and Natural Heritage. Technical Report november, UNESCO, Paris.

UNESCO, 2008. Operational guidelines for the implementation of the world heritage convention. Technical Report July, UNESCO World Heritage Centre.

Wallach, H. M., 2006. Topic modeling: beyond bag-of-words. *Proceedings of the 23rd international conference on Machine learning*, 977–984.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E., 2016. Hierarchical attention networks for document classification. *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.

Zhang, F., Zhou, B., Ratti, C., Liu, Y., 2019. Discovering place-informative scenes and objects using social media photos. *Royal Society open science*, 6(3), 181375.