A Semantically Annotated JSON Metadata Structure For Open Linked Cultural Data In Neo4j

Georgios Drakopoulos
Institute of Informatics and Telecommunications
NCSR "Demokritos"
gdrakop@iit.demokritos.gr

Yorghos Voutos
Department of Informatics
Ionian University
c16vout@ionio.gr

Evaggelos Spyrou
Institute of Informatics and Telecommunications
NCSR "Demokritos"
espyrou@iit.demokritos.gr

Phivos Mylonas
Department of Informatics
Ionian University
fmylonas@ionio.gr

ABSTRACT

Digital culture is a mainstay of the emerging 6V era as both the digitization of existing cultural data and the creation of original native digital cultural content directly lead to the generation of large data volumes which may well be bursty, unstructured or semi-structured, and inherently multimodal. One way to address the increased complexity associated with 6V data is to create metadata structures which succinctly summarize a segment of the underlying cultural data, add semantic information, and serve as indexing and clustering points. This conference paper proposes a JSON description for cultural items which is tailored to the needs of the digital culture domain, generic enough to summarize the vast majority of cultural items, and flexible enough to be extended should the need arise. Moreover, it supports by construction multilevel clustering, semantic annotations, and links to and from relevant cultural items. As a concrete example, the proposed description has been imported to an instance of a Neo4j graph database. The latter takes full advantage of the capabilities offered by the proposed JSON metadata structure, especially the dynamic directed linking between similar cultural items. The abovementioned Neo4j is populated with cultural items from the Ionian Islands, a Greek region with rich cultural tradition.

CCS CONCEPTS

• **Information systems → Ontologies**; *Version management*; *Document topic models*; • **Applied computing → Arts and humanities**; **Document searching**.

KEYWORDS

digital culture, cultural items, multilevel clustering, similarity metrics, cultural metadata, metadata structures, interoperability requirements, JSON, linked open data, higher order analytics, Neo4j

PCI '19, November 28 – 30, 2019, Nicosia, Cyprus

ACM ISBN 1-978-xxxxx…$15.00
https://doi.org/10.xxxx/xxxx

ACM Reference Format:
Georgios Drakopoulos, Yorghos Voutos, Evaggelos Spyrou, and Phivos Mylonas. 2019. A Semantically Annotated JSON Metadata Structure For Open Linked Cultural Data In Neo4j. In *PCI '19, November 28 – 30, 2019, Nicosia, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.xxxx/xxxx

1 INTRODUCTION

Culture has diachronically been the primary way of collective expression in human societies. Questions of cultural influence date at least as back as the Babylonian empire and its subjugates, go through the relationship between the Classical Greek city-states and the Roman empire, and meet the Chinese Mandarins. As the digital age rapidly progresses and enters the so-called 6V phase, the need for managing both historical and new cultural content or other culture-related objects collectively called *cultural items* including the generation of original material, the preservation of existing one, and the efficient online retrieval gradually becomes apparent. In fact, the online almost insatiable quest for at least new pop cultural content may well surpass even the predictions of Andy Warhol himself. Recent indications of the demand for cultural items include the huge success of social media, of cinematic universes with very long story arcs, and of lengthy book series. Under this view, the success of the film *Ready player one*[1] which relied very heavily on pop culture references takes on a new meaning.

However, managing digital or digitized cultural content is by no means a trivial task. On the contrary, only the collection and cataloguing of the various watermarked, commented, or otherwise modified or processed copies and variants of the same cultural item as well as any recovered physical or online ghost copies thereof poses a significant algorithmic and ontological challenge.

To address these issues a plethora of metadata templates created to meet the needs of various community groups has been developed. So now exist metadata templates for bibliographic evidence, archival material, museum objects, and abstract concepts and ideas just to name a few. Along with these metadata ensuring semantic interoperability come platforms that process digital cultural objects.

The threefold primary research objective of this conference paper consists in the following:

[1] https://www.imdb.com/title/tt1677720/

- The proposal of a JSON metadata description which is versatile enough to represent a broad spectrum of cultural items.
- The development of similarity metrics and analytics based on set theory and probability theory respectively which take advantage of the abovementioned metadata structure.
- The demonstration of the potential of both the metadata and the analytics by implementing them on an instance of a Neo4j database populated with cultural items from the region of Ionian islands.

The remainder of this work is structured as follows. Section 2 briefly reviews the scientific literature about cultural content management systems and cultural metadata. The epicenter of section 3 is the proposed JSON metadata for representing cultural content. Section 4 demonstrates how analytics can be implemented in Neo4j. The results from the test system are explained in section 5. Section 6 recapitulates the main points of this conference paper and paves the way for future research direction. Acronyms are explained the first time they are encountered in the text with the sole exception of names of organizations or corporations. Finally, the notation of this conference paper is summarized in table 1.

Table 1: Notation of this conference paper

Symbol	Meaning
$\triangleq$	Definition or equality by definition
$\{s_1, \ldots, s_n\}$	Set with elements $s_1, \ldots, s_n$
$(t_1, \ldots, t_n)$	Tuple with elements $t_1, \ldots, t_n$
$\lvert S \rvert$	Set or tuple cardinality
$S_1 \setminus S_2$	Asymmetric set difference
ν_{S_1,S_2}	Asymmetric Tversky set similarity index
τ_{S_1,S_2}	Tanimoto set similarity metric
ϑ_{S_1,S_2}	Sørensen set similarity metric
$\mathrm{E}[X]$	Mean value of the random variable X
$\mathrm{Var}[X]$	Variance of the random variable X

2 PREVIOUS WORK

The importance of metadata for cultural items has been extensively studied in the relevant scientific literature from a number of aspects. Online cultural databases such as Google Arts and Culture[2] -formerly the Google Art Project- promote the idea of preserving cultural heritage.

In [15] the founding principles of metadata as well as their limits are explored. Along a similar reasoning [3] proposes management practices for cultural metadata repository and [19] and [36] for cultural portals. Automated and user-annotated metadata generation techniques are presented respectively in [20] and in [43]. Cultural metadata can also play a crucial role in bridging differences between cultures which otherwise would hinder the study of a cultural item as shown in [31]. The connection between cultural data and sentiment analysis is investigated in [47].

Specific applications of cultural metadata include the semantic annotations for WWII-related metadata are given in [25]. Moreover, tools for preserving cultural items in 3D formats are discussed in [16] and also in [30]. A conceptual reference model for cultural metadata is presented in [18], whereas the applications of the OAI protocol to the management of cultural repositories are explored in [39]. Higher order ontologies such as those studied in [11] can be applied to the cultural domain as well. The notion of generating recommendations based on metadata among similar artistic subjects and individual artists has been studied in [33] and in [5], where a method based on acquiring, filtering, and condensing metadata related to cultural items was developed, and in [1] where a bottom-up approach for recommendation engines for artistic similarity is presented.

The advent of NoSQL databases signifies a shift towards specialized processing of data types which are not of tabular nature. The four major technological branches of NoSQL databases are shown in table 2.

Neo4j is a prominent graph database [44][32] and a major driver behind the emerging Graph Query Language (GQL) standard[3]. Moreover, it supports a high level, declarative, ASCII art query language named *Cypher* - an obvious reference to *The Matrix*[4] which allows pattern- and constraint-based search of the graph. The patterns can be combinatorial, such as paths, triangles, and neighborhoods, as well as semantic ones, including edge and vertex labels. The basic structure of a graph query is the following:

```
[with <pattern> as <pattern>]
match <pattern>
[where <constraints>]
return <expression>
[order by <criterion> [desc]]
```

The notion of developing a database as a tool for cultural heritage information management was presented by Kioussi et. al [21]. More specifically, the authors described an innovative semantic based knowledge for managing conservation interventions related to the protection of cultural heritage buildings. They've focused mainly on the development of a Non-SQL relational database following the according NDT protocols. Their work permits elaboration among several data, especially the NDT type, in order to analyze, diagnose and evaluate the effectiveness of conservation materials through nondestructive testing standard investigation. Kioussi et al. created a dataset consisting of real case scenarios, which allow further involvemnet of ontological schemes for complex relationships of paramount importance for cultural heritage applications. The clustering of linked cultural items is by no means a trivial challenge since typically they are multimodal and multidimensional. To this end graph partitioning methods such as those proposed in [12] or graph resilience techniques as the one found in [14] can be employed. Finally, in the special case of graph database complex and nested JSON structures can be serialized with the engine proposed in [45].

3 JSON METADATA

JavaScript Object Notation (JSON) metadata description has been formally introduced in RFC 7159[5] and since then it is one of the

[2] https://artsandculture.google.com/

[3] https://neo4j.com/press-releases/query-language-graph-databases-international-standard/

[4] https://www.imdb.com/title/tt0133093/

[5] https://tools.ietf.org/html/rfc7159

Table 2: NoSQL technologies

Database	Data type	Standard	Software
Graph	Linked data	Property graph, JSON-LD, RDF	Neo4j, TitanDB
Key-value	Associative array	JSON, YAML, XML	Redis
Column family	Structured columns	JSON, BSON	Cassandra, HBase
Document	Structured document	JSON	MongoDB

principal ways to describe human-readable data. In fact, its popularity has lead to the implementation of highly efficient JSON parsers for a plethora of programming languages including C++, Java, Rust, and Clojure. In Python a JSON parser is part of the standard library and can be invoked with just the following command:

```
import json
```

Recently, because of the increased significance of graph databases both in research and in production-grade enterprise environments, a JSON-based description especially tailored to linked data which may well be multiply nested termed *JSON for Linked Data* (JSON-LD)[6] has been developed by the W3 consortium [41][46] and has been successfully applied to a number of scenarios [27][28].

An alternative to JSON and JSON-LD are the ubiquitous Resource Description Framework (RDF) triplets [10][2]:

$$(\text{subject}, \text{predicate}, \text{object}) \tag{1}$$

The above carry semantic annotation which is strictly enforced, making RDF triplets and ideal axiom representation form for automated reasoners like HermiT[7] and Pellet[8].

At this point and before examining the fields of the proposed JSON metadata structure two important issues should be mentioned:

- RDF and JSON-LD work naturally with graph databases. Yet, for interoperability purposes with other database types the original JSON format has been selected.
- Complex cultural items can essentially be broken down to rudimentary cultural elements termed *memes* [7], which in turn implies that they can be digitized to a varying yet satisfying extent. Therefore, evolutionary models for discrete systems may -but not necessarily- be applied to complex cultural environments [9].

Figure 1 shows the fields of the proposed JSON metadata structure for cultural objects. These are tailored specifically to succinctly describe complex, multimodal, and multidimensional cultural objects which may well have multiple versions. For instance, a piece of literature may be translated in various languages other than the original one by more than one translators. This section describes the most important of the fields, whereas the meaning of each field is also mentioned summarily in table 3.

The *metadata* field contains the *version* subfield with the current metadata structure version and the *fields* subfield which has the field names of the JSON structure. These self-referential fields allow the coexistence in the same database instance of different versions

```
{
  "metadata": {
    "version": 123 ,
    "fields": ["field1", "field2"]
  } ,
  "mgt_info": {
    "id": {
      "primary": 123 ,
      "secondary": 456
    } ,
    "type": "document" ,
    "description": "info about the document" ,
    "authors": ["author1", "author2"] ,
    "curators": ["curator1", "curator2"] ,
    "creation_date": 000000 ,
    "modalities": ["modality1", "modality2"] ,
    "location": {
      "name": "name" ,
      "uri": "uri
    } ,
    "digital_rights": ["DRM1", "DRM2"] ,
    "reference_to_id": [001, 002] ,
    "reference_from_id": [003, 004, 005]
  } ,
  "work_info": {
    "lost": false ,
    "disputed": false ,
    "authors": ["author1", "author2"] ,
    "locations": {
      "names": ["name1", "name2"] ,
      "uri": ["uri1", "uri2"]
    } ,
    "origin": "document origins" ,
    "time": "timestamp" ,
    "notes": "special notes" ,
    "keywords": ["keyword1", "keyword2"] ,
    "scholars": ["scholar1", "scholar2"] ,
    "language": "language"
  }
}
```

Figure 1: Proposed JSON metadata description.

of metadata as well as the easy discovery of fields from database applications, provided that the latter are aware of the semantics of each version. In any case, at least a limited interoperability between an application and the JSON description of different versions can be achieved. Moreover, the inclusion of these fields are mandated by the best practices of JSON.

The *digital rights management* field is an important addition to the proposed metadata representation since quite often the release of cultural items is tied to specific digital rights agreements which may or may not allow actions such as free redistribution or modification. Especially currently released digital content in various online multimedia platforms such as YouTube or Vimeo is protected by strict intellectual property regulations [37][17].

The *disputed work* field is a Boolean value which indicates whether there is sufficient evidence that a certain cultural item may be falsely attributed to its putative creator [4]. This field covers cases such as that of pseudo-Xenophon or pseudo-Apollodorus.

[6] https://json-ld.org/
[7] http://www.hermit-reasoner.com/
[8] https://github.com/stardog-union/pellet

The *modalities* field cover the case where a cultural item can be multimodal. For instance, a theatrical play may well be accompanied by original music score or costumes. Thus, it is an array in order to accommodate this case. The permitted modalities currently are:

- "text"
- "audio"
- "video"
- "picture"

The *languages* field is an array denoting the number of languages the current cultural item contains or has been translated to, including local dialects. This is especially important for regions such as the Ionian islands which have a distinctive local dialect with strong Italian and French influences.

The *document type* field describes the type of the cultural item. Currently, the following values are supported:

- "literature"
- "theater"
- "film"
- "painting"
- "music"
- "food"
- "clothing"
- "sculpture"

The *keywords* field has keywords pertaining to the current cultural items. In contrast to ordinary terms, keywords have increased semantic importance and capture essential aspects of the cultural item.

The *authors* field refers to the creators of the current cultural item, which can be unknown, anonymous, or a collective entity, as for instance happens with folk music.

The fields *reference_to_id* and *reference_from_id* contain the cultural items which are pointed to by or point to the current cultural item. These fields essentially create the linked nature between the various cultural items and facilitate the metadata storage in Neo4j.

Discovering and representing location mentions in cultural objects can be a challenging task. Relying on a URI instead of geographical coordinates has the following advantages:

- A location may well be a purely imaginary place and yet there may exist a substantial amount of information about it in cultural or scholar works. For instance, *Atlantis* mentioned in Plato's dialogues *Timaeus* and *Critias* has deeply influenced the works of many prominent Western philosophers including Athanasius Kircher, Francis Bacon, and Thomas Moore.
- References can be made to existing yet undiscovered locations with rich history as in the case of ancient Troy before the paramount archaeological discovery of Heinrich Schliemann.
- Finally, the URIs can be naturally linked with the historical and cultural data, which are typically multimodal, associated with a given geographical location in a machine understood way which allows their reliable and quick retrieval.

The advantages offered by the proposed cultural metadata structure are the following:

- Scalability: Instead of handling directly the raw data, operations are executed on the metadata, leading to added efficiency when the raw data size increases. Moreover, Neo4j databases scale up gracefully with the metadata size.
- Flexibility: Neo4j provides an API for developers with advanced graph algorithms. Additional analytics can be built on the client or imported from third party libraries.
- Interoperability: Interoperabiliity is structured into steps that can be followed to enrich and publish the content in the form of linked data. The introduction to the linked data and their principles are based on the processes for producing ontological snapshots from metadata. Furthermore, linked open data (LOD) are enriched and linked into external sources that require further data analysis techniques.

Finally, as stated earlier, table 3 serves as a reference point for the various JSON fields.

4 METADATA QUERIES AND ANALYTICS

4.1 Analytics

Once the database instance has been populated with the metadata, the next step is to develop a number of analytics for discovering latent patterns in the JSON entries.

Assume there are in total c cultural items in total stored in the database with $n < c$ of them being original versions. Moreover, let for the i-th original item be c_i variants. Then:

$$c \triangleq \sum_{i=1}^{n}(1+c_i) = n + \sum_{i=1}^{n} c_i \tag{2}$$

Notice that each of the n original version items introduces an equivalence class C_i with $|C_i| = 1+c_i$ and whose probability among the total items stored in the database is:

$$p_i \triangleq \frac{|C_i|}{\sum_{j=1}^{n}|C_j|} = \frac{1+c_i}{c} = \frac{1+c_i}{n+\sum_{i=1}^{n} c_i} \tag{3}$$

The distribution of p_i is crucial for characterizing the particular instance of the database. Some of the most common measures are:

Definition 4.1 (Dominant item). The dominant item is the one whose probability is the maximum among those in the distribution:

$$p^* \triangleq \max_{1\le i\le n} p_i \tag{4}$$

An important property of the dominant class is that:

$$p^* > \frac{1}{n} \tag{5}$$

Proof. Assume that for each i, $1 \le i \le n$ it holds that:

$$p_i < \frac{1}{n} \tag{6}$$

Then, for the sum of each p_i it holds that:

$$\sum_{i=1}^{n} p_i < \sum_{i=1}^{n} \frac{1}{n} = 1 \tag{7}$$

This clearly violates the definition of probability distribution. □

Table 3: JSON fields

Field	Type	Meaning
metadata.version	Integer	Version number of the particular JSON description
metadata.fields	Array	Name of the fields in this particular description version
mgt_info.id.primary	Integer	Unique identifier for the main version of the specific cultural item
mgt_info.id.secondary	Integer	Unique identifier for each variant of the main version
mgt_info.type	String	Type of the item
mgt_info.description	String	Description of the item
mgt_info.authors	Array	Name(s) of item author(s)
mgt_info.curators	Array	Name(s) of curator(s)
mgt_info.creation_date	Integer	Cultural item creation date
mgt_info.modalities	Array	Names of any modalities
mgt_info.location.name	String	Name of the location associated with the item
mgt_info.location.uri	String	URI about the location associated with the item
mgt_info.digital_rights	Array	The various DRMs covering the particular cultural item
mgt_info.reference_to_id	Array	Similar cultural items which point to the current one
mgt_info.reference_from_id	Array	Similar cultural items the current one points to
mgt_info.origin	String	Notes about the origin of the cultural item
mgt_info.time	String	Era of the cultural item
mgt_info.notes	String	Notes about the cultural items
mgt_info.keywords	Array	Keywords relevant to the cultural item
mgt_info.scholars	Array	List of scholars who have worked on the particular cultural item
mgt_info.languages	Array	List of languages the current cultural item has been translated to

Another probability mass function which can reveal important information about the stored metadata is the distribution of document types, as defined by the *mgt_info.type* JSON field. This distribution is an instrumental factor in query design an optimization. Let D_j denote the j-th document class with $|D_j|$ documents belonging to that class. Then the probability of the j-th document class among the m classes is:

$$q_j \triangleq \frac{|D_j|}{\sum_{k=1}^{m} |D_k|} = \frac{d_j}{\sum_{k=1}^{m} d_k} \tag{8}$$

The dominant item class is defined as follows:

Definition 4.2 (Dominant class). The dominant class is the one whose probability is the maximum among those in the distribution:

$$q^* \triangleq \max_{1 \le j \le m} q_j \tag{9}$$

It is safe to assume that $m \gg n$ since in a mature database instance each item class should contain at least one -but usually much more- item plus its variants.

Categorical data have been mapped through the one-hot encoding technique to binary vectors in order to convert them to numerical values.

Notice that unless the random variable describing the cultural item types X or the number of variants of each cultural item Y is normal or Gaussian or in general belongs to the exponential family of distributions, then the knowledge of the first moment, i.e. the mean value $\mathrm{E}[X]$, and the second central moment, i.e. the variance $\mathrm{Var}[X]$, alone is insufficient to describe the distribution. In order to remedy this, the generating function of the distribution should be considered:

$$\varphi_X(z) \triangleq \mathrm{E}\left[z^X\right] = \sum_{i=1}^{n} p_i z^{c_i} \tag{10}$$

In a similar manner the generating function of the document type distribution $\varphi_Y(z)$ can be defined. Notice that since the number of cultural items n as well as the the number of document classes m are finite, then there are no convergence issues in either case.

Observe that the use of a higher order indicator such as the generating function is, intuitively at least, appropriate for linked cultural items. Since the latter are represented by a graph, a combinatorial object which is inherently distributed and can naturally express higher order relationships through its paths, it makes sense to employ an analytic tool like $\varphi(\cdot)$ which is designed to take such relationships into account.

Discovering similarities not only between variants of the same cultural item but also across classes of items is of paramount importance in the cultural preservation field since it can lead to the discovery of latent patterns in cultural items, which in turn may reveal a general trend.

Given that finding ground truth interpretations for cultural items may not be feasible, especially those coming from eras which are not adequately historically documented, a number of non-supervised learning techniques should be employed in order to cluster items.

Self-organizing maps (SOMs) constitute a class of unsupervised learning networks relying on a combination of a relaxed Hebbian learning rule and of an optional dormant neuron correction mechanism through a bias [22][24][23]. Alternatively, similarity trees have been proposed in [38] in order to classify two objects u and v according to their similarity. The latter is expressed by the length

of the path connecting the end nodes of the tree which correspond u and v. Although both approaches scale relatively well with the input size, simpler similarity metrics were chosen.

For similarity comparison between metadata corresponding to original cultural item versions the Tanimoto metric [29][26] shown in (11) can be used.

$$\tau_{S_1,S_2} \triangleq \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = \frac{|S_1 \cap S_2|}{|S_1| + |S_2| - |S_1 \cap S_2|} \tag{11}$$

The second form of the Tanimoto coefficient is more efficient for large databases since intersection queries are typically optimized, whereas the original set cardinalities can be efficiently approximated by estimators [13][8].

An alternative to equation (11) is the Sørensen set similarity index which is defined as follows:

$$\vartheta_{S_1,S_2} \triangleq 2\frac{|S_1 \cap S_2|}{|S_1| + |S_2|} \tag{12}$$

Notice that both equations (11) and (12) are symmetric in the sense that S_1 and S_2 are interchangeable. As a rule, this is not desirable when comparing an original version with its variant, since the former is a template for the latter. In this case, the Tversky asymmetric index [42][35] defined in equation (13) can be used.

$$\nu_{S_1,S_2} \triangleq \frac{|S_1 \cap S_2|}{\alpha_0 |S_1 \setminus S_2| + (1-\alpha_0)|S_2 \setminus S_1| + |S_1 \cap S_2|} \tag{13}$$

In equation (13) the parameter α_0 is bounded as follows:

$$0 \leq \alpha_0 \leq 1 \tag{14}$$

The average Tanimoto I similarity between item classes D_{j_1} and D_{j_2} can be defined as the arithmetic mean of the pairwise comparisons between all original cultural items of the respective classes:

$$\tau_{D_{j_1},D_{j_2}} \triangleq \frac{1}{(d_{j_1}-1)(d_{j_2}-1)} \sum_{I_{k_1} \in D_{j_1}} \sum_{I_{k_2} \in D_{j_2}} \tau_{I_{k_1},I_{k_2}} \tag{15}$$

In (15) the Tanimoto similarity is computed over the selected metadata fields for each pair of distinct cultural items I_{k_1} and I_{k_2} where I_{k_1} belongs to D_{j_1} and I_{k_2} belongs to D_{j_2}. Also, recall that $d_{j_1} = \left|D_{k_1}\right|$ and $d_{j_2} = \left|D_{j_2}\right|$.

Similarly, the average Tanimoto II similarity between item classes D_{j_1} and D_{j_2} is defined by taking into account all cultural items including variants.

At this point it should be highlighted that the above essentially constitute a multilevel partition or a clustering hierarchy of the entire cultural item graph:

- Item classes are at the most abstract level where m vertices exist along with any edges denoting similarities across item classes.
- Original cultural items are at an intermediate level where there exist n vertices and the corresponding edges denoting similarities between these items.
- Cultural items in general are at the highest granularity level where c vertices exist along with a plethora of edges.

4.2 Queries

A Neo4j instance can be controlled with either commands entered directly in the database console or through a client which also runs a database driver. In the test implementation the particular driver was py2neo[9], installed through pip. This driver interfaces directly with the Python source code. py2neo provides a functional interface where vertices, edges, and their properties can be seen and handled as Python objects or, alternatively, allows strings containing valid Cypher queries or database control commands to be executed with the results returned in a list. The latter option was selected.

In order to insert a new document vertex the following Cypher command can be typed at the Neo4j console or issued through the Python client. For simplicity only one vertex field is shown. The actual source code contains all the JSON fields. Notice that a Neo4j vertex can contain arbitrary data, which may differ depending on the vertex type, in key-value format.

```
merge (u {document: "document"})
return a
```

The insertion of a new edge is achieved with the following Cypher command:

```
match (u),(v)
where u.type = "type1" and v.type = "type2"
create (u)-[r:RELTYPE]->(v)
return r
```

Currently the following three edge labels are supported, as also shown in figure 2:

- VARIANT: Edges of this type connect vertices representing original versions with variants.
- REFERENCES: If a cultural item references somehow, usually as a text, another item, then they are connected with this edge type.
- SIMILAR: Used only between vertices representing original vertices, it appears only when the Tanimoto coefficient exceeds a certain threshold. Also, the similarity can be defined on a number of fields such as the authors, the curators, or the the keywords. This allows various queries to be created.

5 RESULTS

Ionian islands form a Greek region which is especially renowned from its rich cultural tradition since at least the classical antiquity [34] and which is still present and ongoing today [40][6].

Although Neo4j is schema-less in the traditional relational sense, it is still possible to see all the edge and vertex labels as well as the possible connection pattern between them. Figure 2 can be generated by typing at the database console:

```
:call db.schema()
```

Table 4 contains a summary of the test dataset which serves as a demonstration tool for the capabilities of the proposed metadata structure. This dataset consists of three classes of cultural items from the Ionian islands tradition, namely poems (D_1), musical excerpts (D_3), and theatrical plays (D_2). These cultural items belong to the same century.

[9] https://py2neo.org/v4/

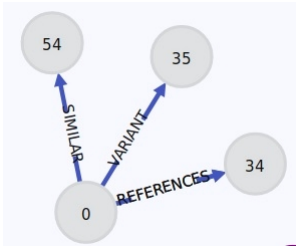

Figure 2: Neo4j schema.

Table 4: Dataset summary

Item	Variants	Item	Variants	Item	Variants
Poem#1	1	Play#1	1	Play#9	1
Poem#2	2	Play#2	1	Play#10	3
Poem#3	5	Play#3	2	Play#11	2
Poem#4	1	Play#4	1	Music#1	1
Poem#5	2	Play#5	1	Music#2	4
Poem#6	1	Play#6	3	Music#3	3
Poem#7	1	Play#7	2	Music#4	2
Poem#8	3	Play#8	1	Music#5	1

Using the notation developed earlier for the specific dataset it follows that $m = 3$, $n = 24$, and $c = 46$. Moreover:

$$\begin{aligned}(d_1, d_2, d_3) &= (8, 11, 5)\\ (q_1, q_2, q_3) &= \left(\frac{8}{24}, \frac{11}{24}, \frac{5}{24}\right)\end{aligned} \tag{16}$$

Thus $q^* = q_2$.

Focusing on the *keyword* set of the JSON metadata, the average Tanimoto I and II similarity coefficients among the three item classes are shown in table 5 and 6 respectively.

Table 5: Average Tanimoto I (keywords)

Class	D_1	D_2	D_3
D_1	1	0.3117	0.1733
D_2	0.3117	1	0.2114
D_3	0.1733	0.2114	1

Table 6: Average Tanimoto II (keywords)

Class	D_1	D_2	D_3
D_1	1	0.3542	0.1983
D_2	0.3542	1	0.2333
D_3	0.1983	0.2333	1

From tables 5 and 6 the following can be deduced:

- The values for Tanimoto II similarity coefficient similarity are bigger than the corresponding ones of the Tanimoto I. This can be attributed to the large number of items which have similar keywords.
- There appears to be a stronger connection between D_1 and D_2 and between D_2 and D_3 than between D_1 and D_3. This may be explained from the fact that theatrical play writers also wrote poems and that many theatrical plays of the Ionian islands rely on music. On the other hand, typically the number of poems which are turned to songs is low. Thus, a weaker connection between music and poetry is expected.
- The values of both coefficients are relatively low, suggesting that all three categories can be partitioned based on the keywords criterion.

Along a similar line of reasoning, computing the same two similarity coefficients for the *authors* JSON field yields tables 7 and 8.

Table 7: Average Tanimoto I (authors)

Class	D_1	D_2	D_3
D_1	1	0.2489	0.1366
D_2	0.2489	1	0.1875
D_3	0.1366	0.1875	1

Table 8: Average Tanimoto II (authors)

Class	D_1	D_2	D_3
D_1	1	0.2666	0.1805
D_2	0.2666	1	0.2250
D_3	0.1805	0.2250	1

From tables 7 and 7 the following can be concluded:

- The patterns from tables 5 and 6 are repeated. This can be explained as the author coherence can lead to keyword coherence.
- The values of the author tables are clearly lower than the corresponding in the keyword tables. This can be attributed to the fact that the keyword sets typically have bigger cardinalities compared to the author ones as to a given author more than one keywords may well apply.

6 CONCLUSIONS AND FUTURE WORK

This conference paper presented a JSON metadata structure for cultural items. Moreover, it discussed a number of analytics based on set theory and higher order probability theory. For demonstration purposes the latter have implemented in Python with the JSON metadata stored in a Neo4j instance. The test dataset was constructed from cultural items from the region of Ionian islands, a Greek region renowned for its cultural heritage.

The above become crucial since many cultural institutions explicitly mention in their mission statement that they aim at making their collections available to the general public and the Web presents a nearly unique opportunity to achieve this. To seize that opportunity, they are often willing to take reasonable risks associated with making their materials available on the Web.

The work presented in this conference paper can be naturally extended in a number of ways. First and foremost, in order to deduce useful historical information about cultural evolution spatial and temporal windows should be added to the metadata structure. This can be accomplished by the introduction of spatial data structures such as Quad trees and R trees. For instance, links to local or global historical events can explain the history of institutions, mainly of political and educational, of the Ionian islands. Moreover, the inclusion of numerical fields would also be an important extension to the proposed metadata structure. In this case, appropriate analytics should be developed in order to take into account the new fields. Finally, the clustering hierarchy can be extended to four or more levels. However, this poses the question of how many levels there should be since the a sixth or seventh layer variant may have little or no meaning.

ACKNOWLEDGMENTS

This research has been co-financed by the European Union and Greek national funds through the Competitiveness, Entrepreneurship and Innovation Operational Programme, under the Call "Research – Create – Innovate", project title: "Development of technologies and methods for cultural inventory data interoperability", project code: T1EDK-01728, MIS code: 5030954.

REFERENCES

[1] Shinhyun Ahn and Chung-Kon Shi. 2009. Exploring movie recommendation system using cultural metadata. In *Transactions on Edutainment II*. Springer, 119–134.
[2] Grigoris Antoniou and Frank Van Harmelen. 2004. *A semantic Web primer*. MIT Press.
[3] Murtha Baca. 2003. Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information. *Cataloging & classification quarterly* 36, 3-4 (2003), 47–55.
[4] Chris Barker. 2003. *Cultural studies: Theory and practice*. SAGE.
[5] Stephan Baumann and Oliver Hummel. 2003. Using cultural metadata for artist recommendations. In *International Conference on Web Delivering of Music*. IEEE, 138–141.
[6] Sylvia Benton. 1932. The Ionian Islands. *Annual of the British School at Athens* 32 (1932), 213–246.
[7] Susan Blackmore. 2000. *The meme machine*. Vol. 25. Oxford Paperbacks.
[8] Abhinandan Das, Sumit Ganguly, Minos Garofalakis, and Rajeev Rastogi. 2004. Distributed set-expression cardinality estimation. In *VLDB*. VLDB Endowment, 312–323.
[9] Richard Dawkins. 1981. In defence of selfish genes. *Philosophy* 56, 218 (1981), 556–573.
[10] Stefan Decker, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, and Ian Horrocks. 2000. The semantic Web: The roles of XML and RDF. *IEEE Internet Computing* 4, 5 (2000), 63–73.
[11] Georgios Drakopoulos, Andreas Kanavos, Phivos Mylonas, Spyros Sioutas, and Dimitrios Tsolis. 2017. Towards a framework for tensor ontologies over Neo4j: Representations and operations. In *IISA*. IEEE, 1–6. https://doi.org/10.1109/IISA.2017.8316441
[12] Georgios Drakopoulos, Andreas Kanavos, and Konstantinos Tsakalidis. 2017. Fuzzy Random Walkers with Second Order Bounds: An Asymmetric Analysis. *Algorithms* 10, 2 (2017). https://doi.org/10.3390/a10020040
[13] Georgios Drakopoulos, Stavros Kontopoulos, and Christos Makris. 2016. Eventually consistent cardinality estimation with applications in biodata mining. In *SAC*. ACM.
[14] Georgios Drakopoulos, Xenophon Liapakis, Giannis Tzimas, and Phivos Mylonas. 2018. A Graph Resilience Metric Based On Paths: Higher Order Analytics With GPU. In *ICTAI*. IEEE, 884–891. https://doi.org/10.1109/ICTAI.2018.00138
[15] Erik Duval, Wayne Hodgins, Stuart Sutton, and Stuart L Weibel. 2002. Metadata principles and practicalities. *D-lib Magazine* 8, 4 (2002), 1082–9873.
[16] Achille Felicetti and Matteo Lorenzini. 2011. Metadata and tools for integration and preservation of cultural heritage 3D information. *Geoinformatics FCE CTU* 6 (2011), 118–124.
[17] Nic Garnett. 2001. Digital rights management, copyright, and Napster. *ACM SIGCOM Exchanges* 2, 2 (2001), 1–5.
[18] Manolis Gergatsoulis, Lina Bountouri, Panorea Gaitanou, and Christos Papatheodorou. 2010. Mapping cultural metadata schemas to CIDOC conceptual reference model. In *Hellenic Conference on Artificial Intelligence*. Springer, 321–326.
[19] Eero Hyvönen. 2009. Semantic portals for cultural heritage. In *Handbook on ontologies*. Springer, 757–778.
[20] Krassimira Ivanova, Milena Dobreva, Peter Stanchev, and George Totkov. 2012. *Access to Digital Cultural Heritage: Innovative Applications of Automated Metadata Generation*. Plovdiv University Publishing House "Paisii Hilendarski".
[21] Anastasia Kioussi, Maria Karoglou, Anastasios Doulamis, Christodoulos Fragkoudakis, Petros Potikas, Eftychios Protopapadakis, Ekaterini Delegou, Emmanouil Alexakis, and Antonia Moropoulou. 2019. A Knowledge Json-Based Database for Integrating Multiple Disciplines in Cultural Heritage. In *Nondestructive Evaluation and Monitoring Technologies, Documentation, Diagnosis and Preservation of Cultural Heritage*, Ahmad Osman and Antonia Moropoulou (Eds.). Springer International Publishing, Cham, 121–132.
[22] Teuvo Kohonen. 1990. The self-organizing map. *Proceedings of the IEEE* 78, 9 (1990), 1464–1480.
[23] Teuvo Kohonen. 1997. Exploration of very large databases by self-organizing maps. In *ICNN*, Vol. 1. IEEE, PL1–PL6.
[24] Teuvo Kohonen, Erkki Oja, Olli Simula, Ari Visa, and Jari Kangas. 1996. Engineering applications of the self-organizing map. *Proceedings of the IEEE* 84, 10 (1996), 1358–1384.
[25] Alexander Koplenig. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets –Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities* 32, 1 (2017), 169–188.
[26] Thomas G Kristensen. 2010. Transforming Tanimoto queries on real valued vectors to range queries in Euclidian space. *Journal of Mathematical Chemistry* 48, 2 (2010), 287–289.
[27] Markus Lanthaler and Christian Gütl. 2012. On using JSON-LD to create evolvable RESTful services. In *Third international workshop on RESTful design*. ACM, 25–32.
[28] Markus Lanthaler and Christian Gütl. 2013. Model your application domain, not your JSON structures. In *International conference on World Wide Web*. ACM, 1415–1420.
[29] Alan H Lipkus. 1999. A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry* 26, 1-3 (1999), 263–265.
[30] Emmanuel Maravelakis, Antonios Konstantaras, Athina Kritsotaki, Dimitrios Angelakis, and Michael Xinogalos. 2013. Analysing user needs for a unified 3D metadata recording and exploitation of cultural heritage monuments system. In *International Symposium on Visual Computing*. Springer, 138–147.
[31] Richard McDermott and Carla O'Dell. 2001. Overcoming cultural barriers to sharing knowledge. *Journal of knowledge management* 5, 1 (2001), 76–85.
[32] Justin J Miller. 2013. Graph database applications and concepts with Neo4j. In *Southern association for information systems conference*, Vol. 2324.
[33] Francois Pachet. 2005. Knowledge management and musical metadata. *Idea Group* (2005), 12.
[34] Jim Potts. 2010. *The Ionian Islands and Epirus: A Cultural History*. Oxford University Press.
[35] Michael M Richter. 1993. Classification and learning of similarity measures. In *Information and Classification*. Springer, 323–334.
[36] P Ronzino, S Hermon, and F Niccolucci. 2012. A metadata schema for cultural heritage documentation. *Electronic Imaging and the Visual Arts* (2012), 36–41.
[37] Bill Rosenblatt, Bill Trippe, Stephen Mooney, et al. 2002. Digital Rights Management. *New York* (2002).
[38] Shmuel Sattath and Amos Tversky. 1977. Additive similarity trees. *Psychometrika* 42, 3 (1977), 319–345.
[39] Sarah L Shreeves, Joanne S Kaczmarek, and Timothy W Cole. 2003. Harvesting cultural heritage metadata using the OAI protocol. *Library hi tech* 21, 2 (2003), 159–169.
[40] Christina Souyoudzoglou-Haywood. 1999. *The Ionian Islands in the bronze age and early iron age, 3000-800 BC*. Liverpool University Press.
[41] Manu Sporny, Dave Longley, Gregg Kellogg, Markus Lanthaler, and Niklas Lindström. 2014. JSON-LD 1.0. *W3C Recommendation* 16 (2014), 41.
[42] Amos Tversky and Itamar Gati. 1978. Studies of similarity. *Cognition and categorization* 1, 1978 (1978), 79–98.
[43] Seth Van Hooland, Eva Méndez Rodríguez, and Isabelle Boydens. 2011. Between commodification and engagement: On the double-edged impact of user-generated metadata within the cultural heritage sector. *Library trends* 59, 4 (2011), 707–720.
[44] Jim Webber and Ian Robinson. 2018. *A programmatic introduction to Neo4j*. Addison-Wesley Professional.
[45] Jens H Weber. 2017. GRAPE: A graph rewriting and persistence engine. In *International conference on graph transformation*. Springer, NY, 209–220.
[46] World Wide Web Consortium et al. 2014. JSON-LD 1.0: A JSON-based serialization for linked data. (2014).
[47] Helena Wulff. 2007. *The emotions: A cultural reader*. Berg Publishers, Oxford, UK.