information

MDPI

Article

CIDOC2VEC: Extracting Information from Atomized CIDOC-CRM Humanities Knowledge Graphs

Hassan El-Hajj [1,2,*] **and Matteo Valleriani** [1,2,3,4]

1 Max Planck Institute for the History of Science, 14195 Berlin, Germany; valleriani@mpiwg-berlin.mpg.de
2 BIFOLD—Berlin Institute for the Foundation of Learning and Data, 14195 Berlin, Germany
3 Technische Universität Berlin, 10623 Berlin, Germany
4 Faculty of Humanities, Tel-Aviv University, Tel Aviv-Yafo 69978, Israel
* Correspondence: hhajj@mpiwg-berlin.mpg.de

Abstract: The development of the field of digital humanities in recent years has led to the increased use of knowledge graphs within the community. Many digital humanities projects tend to model their data based on CIDOC-CRM ontology, which offers a wide array of classes appropriate for storing humanities and cultural heritage data. The CIDOC-CRM ontology model leads to a knowledge graph structure in which many entities are often linked to each other through chains of relations, which means that relevant information often lies many hops away from their entities. In this paper, we present a method based on graph walks and text processing to extract entity information and provide semantically relevant embeddings. In the process, we were able to generate similarity recommendations as well as explore their underlying data structure. This approach was then demonstrated on the *Sphaera* Dataset which was modeled according to the CIDOC-CRM data structure.

Keywords: data extraction; knowledge graph; CIDOC-CRM; digital humanities; Sphaera; recommender system

Citation: El-Hajj, H.; Valleriani, M. CIDOC2VEC: Extracting Information from Atomized CIDOC-CRM Humanities Knowledge Graphs. *Information* **2021**, *12*, 503. https://doi.org/10.3390/info12120503

Academic Editors: Pierpaolo Basile and Annalina Caputo

Received: 15 November 2021
Accepted: 30 November 2021
Published: 3 December 2021

1. Introduction

The term knowledge graph was first coined by Google in 2012, defining a large database connecting things, not strings, through different types of relations. Since then, this type of database has played an important role in recommender systems [1], natural language processing and question answering [2], as well as other fields such as the digital humanities [3–8]. Within the humanities, the term knowledge graph can be considered a knowledge organization system [9]. Such a term is rooted in the centuries-old tradition of classifying knowledge, or in the case of libraries, books, based on their meta-data in order to facilitate data retrieval. With the recent developments in information technology, the field of digital humanities has been keen on curating and developing humanities knowledge graphs, taking advantage of openly available knowledge graphs such as DBpedia [10] and Wikidata [11] to enhance their databases. Such knowledge graphs are also being constructed and published according to linked data principles [12]. This increase in knowledge graph use within the digital humanities community opens the door for standardization efforts such as the one presented by the International Committee for Documentation—Conceptual Reference Model (CIDOC-CRM) [13]. This CIDOC-CRM model proposes a theoretical and practical tool for ontology creation in the fields of cultural heritage and humanities, with the aim of creating coherent shareable datasets across multiple institutions. However, the complexity of including all possible heritage and humanities data combinations under one umbrella leads to complex knowledge graph structures. These structures rely on deconstructing knowledge to its simplest generalized atoms (see Figure 1), which are then stored in separate entities, often leading to the creation of long chains of relations connecting the head and tail of relevant entities within a single knowledge graph (see Figure 2).

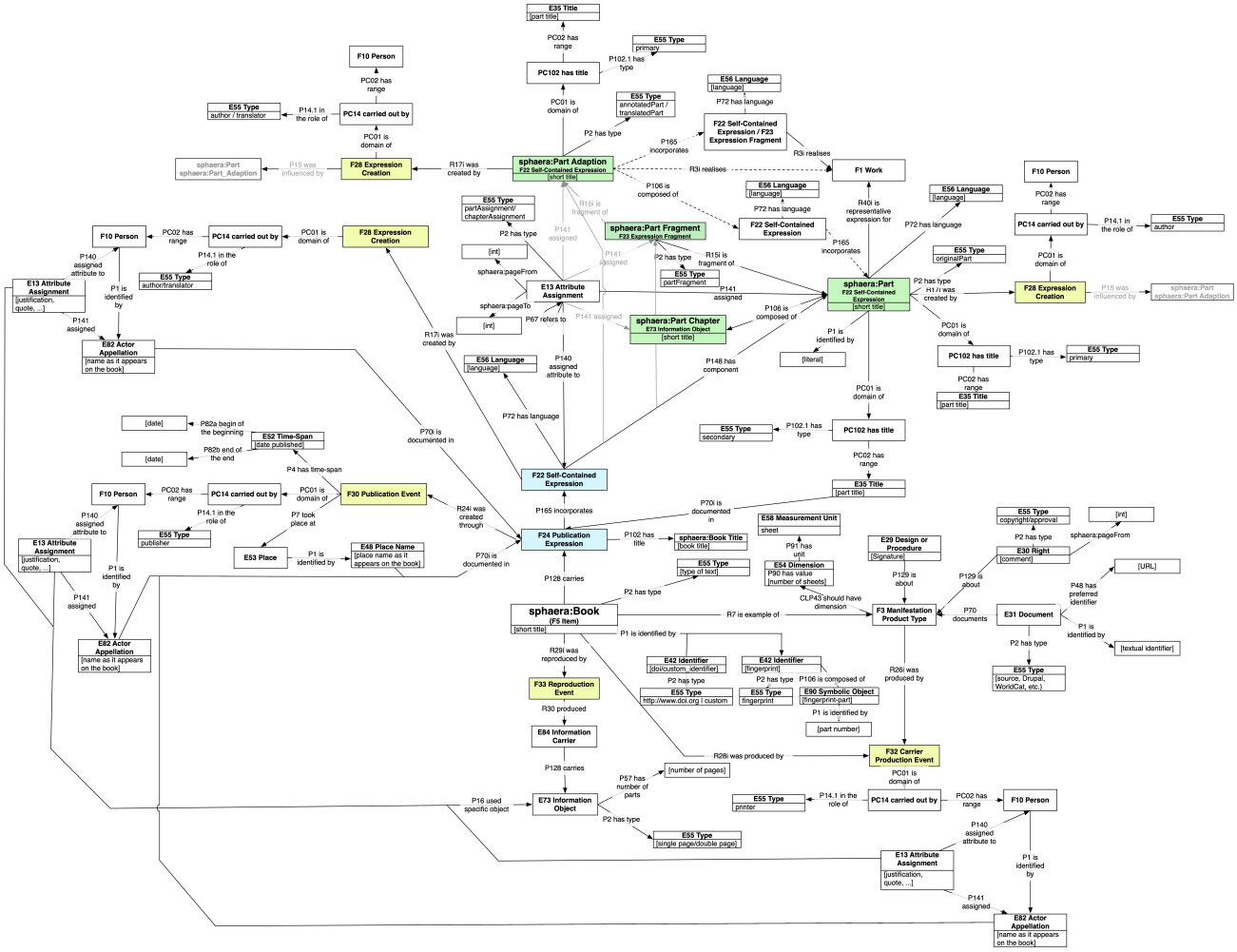

Figure 1. Sphaera CIDOC-CRM data model.

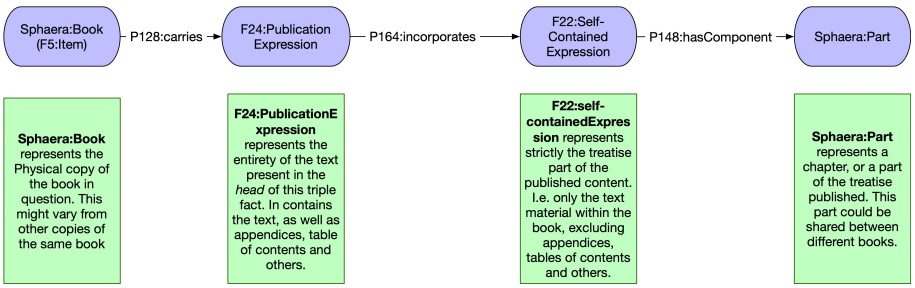

Figure 2. Path expressing the relation between a book and a book part in the *Sphaera* dataset.

In this paper, we present a method that allows researchers working with CIDOC-CRM knowledge graphs to embed certain entities, which we call main entities, in order to better understand the underlying phenomena that these entities encode (e.g., such as the reprint families discussed in Section 6), to better represent these entities, and to better recommend similar objects within the same database, leading to higher public engagement. In order to achieve this, we propose a knowledge graph walk, which we call a relative sentence walk, capable of generating sentence data based on biased random walks through the knowledge graph. These sentences create a representative document of each unique entity, which are subsequently embedded using natural language processing approaches. The importance of this approach is then demonstrated on the *Sphaera* dataset. In the following, we discuss the approaches used in knowledge graph embeddings and the specific needs and limitations

of digital humanities data in Section 2; we then discuss the CIDOC-CRM model which is often used in humanities and cultural heritage databases in Section 3; followed by the *Sphaera* dataset in Section 4; and our method to embed graph entities in Section 5; finally, we present the results in Section 6.

2. Related Work

Knowledge graph embedding is an active field of research, with numerous techniques contributing to the ever-growing field. While we did not aim to present a comprehensive review of knowledge graph embedding in this paper, a short discussion was conducted in order to put into perspective what we aimed to accomplish.

The vast majority of the research conducted on knowledge graph embedding heavily relies on triple sets or facts, which can be noted as (*h, r, t*) where *h* represents the *head*, *r* represents the *relation*, and *t* represents the *tail* of a three-item fact. Such approaches started with translation-based models such as TransE [14] which represents every entity and relation in a low-dimensional vector space and each fact is represented by the connection of *h-r-t* in low dimensional space. The correct entity relation embeddings are then learned through minimizing a hinge ranking loss. While TransE might have been one of the early adopters of translation invariant approaches in knowledge graph embeddings, many subsequent papers built on the TransE model achieved better results by allowing each entity to be represented to its own hyperplane, as in TransH [15], as well as separating the embedding spaces of entities and relations as in TransR [16]. Numerous other translational approaches followed such as TransD, which uses a dynamic mapping matrix and solves the entity diversity by applying a transitional matrix determined by entity relations [17], as well as TransM [18] and TransA [19].

Tensor factorization methods were also used to learn knowledge graph embedding, where the triplet facts are represented as the 3D binary tensor, and each tensor is represented by $X \in \mathbb{R}^{n \times n \times m}$, and n and m represent the number of entities and relations, respectively, [20]. The tensor is populated with binary numbers, where $X_{ijk} = 1$ when the *i*-th entity is related to the *j*-th entity through relation *k*. This approach was used by RESCAL [21], which proposed a rank-d factorization to obtain knowledge graph semantics. Similar tensor factorization approaches were used in methods such as DistMult [22]. However, with the gain in popularity of neural networks, many knowledge graph embedding approaches began to rely on neural networks to learn entity relation embeddings. Semantic Matching Energy (SME) [23] proposes an energy function to judge the authenticity of a triple fact using a shallow neural network. In this method, the embedded triple fact is passed to the network in three separate vectors associated with each component of the triple. The *head* and *tail* of the triple fact are combined, separately, with the *relation* using linear and bilinear functions in two different versions of SME. The parameters of the vector combination functions are then learned during training. Finally, the energy function is computed by matching the left (i.e., *head–relation*) and the right (i.e., *relation–tail*) vectors through a simple dot product. Other approaches, such as ConvKB [24], represent each triplet fact as a three-column matrix where each column represents a part of the fact. This matrix is then fed to a convolutional neural network, where the output vectors are then concatenated and evaluated using a dot product with a weight vector to obtain a score. This scoring is then used to evaluate the authenticity of the triple fact. Neural tensor networks (NTN) [25] are used to calculate the energy score between the *head* and *tail*, $f(h, t)$. NTN replaces the linear layer of a traditional neural network architecture with a bilinear tensor layer. While the above is not a comprehensive review of knowledge graph embedding, it provides an overview of the three main algorithm categories: translation, tensor, and neural network-based approaches.

However, the majority of knowledge graph embedding approaches rely on capturing the relation between triple facts, and often ignore, or do not take into account, the neighborhood information, which can hold very rich information on the nature of the relation between the different entities. Additionally, the totality of knowledge graph embeddings

are suited for traditional knowledge graph architecture, where each entity represents an distinct object, and each relation represents a type of interaction between the said entities. Such approaches are not ideal for peculiar knowledge graph cases, such as the ones often constructed using CIDOC-CRM principles [13], as discussed in the following section.

3. CIDOC-CRM

The International Committee for Documentation—Conceptual Reference Model (CIDOC-CRM) was first introduced in 2006 in order to enable the integration and reconstruction of information from humanities and cultural sources, and to allow for better interpretations of these data [13]. The model itself presents an ontology with the aim of formally defining and structuring the underlying semantics that exist while recording, documenting, and storing heterogeneous data. In doing so, it mainly relies on a predefined set of classes, properties, as well as constraints to ensure consistent modeling between projects and institutions [26]. The main aim is then to ensure the consistent recording of cultural heritage and humanities data, stored in the form of different entities belonging to different pre-set classes, related to each other by structured relations.

Databases constructed based on the CIDOC-CRM approach present a special form of knowledge graph architecture, where class hierarchy defines many of the relations between entities and many of what can be described as attributes are stored in their own entity classes connected to the main entity through descriptive relations. This means that each main entity, while comprehensive in itself, can be completed by a large set of relations that enrich it with descriptive information. For example, an entity describing a book can only be fully understood by following the relevant relations to other descriptive entities, which lead to title entities, page entities, publication entities, as well as other relevant pieces of information. In this case, the CIDOC-CRM allows for the expansion of regular knowledge graph structures to include numerous entity properties, themselves stored in their own entities. Based on this knowledge graph architecture, it follows that a consistent embedding of the main entities should not only look at triple facts, but also take into consideration the main entity's attributes. To better understand the CIDOC-CRM knowledge graph architecture, we present the *Sphaera* knowledge graph [3], modeled according to the CIDOC-CRM principles.

4. The *Sphaera* Knowledge Graph

The project *Sphere: Knowledge System Evolution and the Shared Scientific Indentity of Europe* (sphaera.mpiwg-berlin.mpg.de, accessed on 1 December 2021) aimed to study the mechanisms behind the evolution of knowledge during the early modern period by looking at 359 different editions of European university cosmology textbooks published between 1472 and 1650, centered on the *Tractatus de Sphaera* by Johannes de Sacrobosco (1195–1256) [27]. These editions can be separated into five different categories:

- Original treatises, which are editions that represent the *Tractatus de Sphaera* as a standalone work;
- Annotated original treatises, which are editions that include the *Tractatus de Sphaera* with annotations and commentary by other authors;
- Compilations of texts, which are editions that exclusively include the *Tractatus de Sphaera* among other treatises by different authors;
- Compilations of texts and annotated originals, which include editions that feature the *Tractatus de Sphaera* as the basis for a commentary or annotation, and also includes work by other authors;
- Adaptions, which are editions that contain texts that are heavily influenced by the *Tractatus de Sphaera* in terms of content and structure, but do not include the original text.

Each of the 359 editions within the *Sphaera* corpus is atomized into 450 different textparts, which represent passages of texts covering well-defined subjects within the studied

treatises, as well as paratexts, which are short texts often added to the original content [27]. These text-parts are grouped into five different classes as shown below:

- Content parts, which are the core scientific texts that rarely changed and were considered as the reference text in each edition;
- Paratext—poetry, which represents short poem dedications, often added at the beginning of certain editions;
- Paratext—dedications letters are short passages of dedication, and can often be an indication of the level of prestige of certain editions;
- Paratext—letters to reader or preface, are short passages addressing the reader;
- Paratext—other, which represents a small group of texts that do not fit in the previous three paratext categories.

The entirety of the *Sphaera* corpus is then stored in a knowledge graph originally built by Florian Kräutli [3] using CIDOC-CRM logic [13], as well as its extension, FRBRoo [28]. The central element of this database is the book edition, represented by a single copy of each of each which is considered to be a representative sample for the entire edition print-run. These editions are considered as main entities, which can be described by large networks of property–entity relations, where properties such as content, pages, and other descriptive attributes are stored in their own entity classes, numerous hops away from the book edition entities (see Figure 1). Here, an example is best suited to clarify the logic of such a diagram. Each of the 359 editions within the *Sphaera* knowledge graph is represented by a CIDOC-CRM "F5:Item" Entity Class, as well as the "Sphaera:Book" subclass. This represents the physical copy of the book, and thus any entities related to the "F5:Item" contain information that can describe the physical copy represented by the "F5:Item" in question. For example, the "P43:hasDimension" relation, which connects the "F5:Item" to an "E43:Dimension" entity, aims to express the physical format of the physical book in question. While the "F5:Item" represents the physical copy, its content, i.e., the text contained in the physical book, is expressed by different entities according to the CIDOC-CRM standards, which are connected to the "F5:Item" physical book through a "P128:Carries" relation that connects the "F5:Item" to an "F24:publicationExpression", and this in turn refers to the text intended to be published by the "F5:Item". The text represented by the "F24:publicationExpression" represents the entirety of the book's content, including the page numbers, indices, table of contents and any other additions that might have been added by the publishers to make the book easier to navigate. The content of the treatise in question, without from any addition, is expressed by the "F22:self-containedExpression", related to the "F24:publicationExpression" by a "P165:incorporates" relation. This is to signify that the treatise's text in question is contained within the totality of the book's content, which in itself is related to a "Sphaera:Part" through a "P148:hasComponent" relation, indicating that each treatise is divided into smaller components, or parts, which are often shared across multiple editions within the *Sphaera* corpus (see Figure 2). Other relations along the CIDOC-CRM standards can be seen in Figure 1, while a more detailed look at the logic of CIDOC-CRM and *Sphaera* can be found here [3].

The diagram in Figure 1 forms the base of the *Sphaera* knowledge graph, with each of the 359 editions connected to multiple entities that encode its description. The books themselves are not explicitly connected to each other by any relation. However, books could contain shared parts, be published by the same publisher, or printed by the same printer, or written (or contains written parts) by the same authors; they could also be published in the same city, the same year, as well as a host of implicit connections that one can use to embed the entities close to each other. While historical research added several direct connections between books based on their semantic similarity [27,29], these links were omitted from the analysis for this paper in the aim of only using CIDOC-CRM relations.

It is then based on the desire to extract the underlying similarity from such a complex knowledge graph, and to propose semantically similar books based on the stored data that we propose a general CIDOC2VEC approach applicable to CIDOC-CRM based knowledge graphs and described in the following section.

5. Method: CIDOC2VEC

Due to the complexity of CIDOC-CRM knowledge graphs, the lack of formalized relations between the main entities and the lack of direct triple facts that can express the simple set of properties of a book such as (Book-X, wasAuthoredBy, Person-Y, or Book-X, contains, Part-Y), most of the current approaches presented in Section 2 are inefficient as they are optimized to handle data structures with direct connections between entities. The aim of CIDOC2VEC here is not to predict any missing relation, especially since this is rarely the objective within humanities, but to generate recommendations, or similarity clusters based on the knowledge graph information and intrinsic relations between well-connected main entities with at least a single outgoing relation. Such recommendations could be useful for suggesting similar items within large collections stored on the basis of this data structure, as well as for revealing new information regarding the collection to obtain new insights into the corpus in question. To achieve this goal, we propose CIDOC2VEC, which is composed of two main modules: a relative sentence walk module and natural language processing module.

5.1. Relative Sentence Walk (RSW)

The nature of the syntactical similarity between reading knowledge graphs and natural language is leading to a growing field focused on exploring the symbiosis between natural language processing (NLP) models and knowledge graphs [30–32]; it is this similarity that inspired relative sentence walks (RSWs), and its CIDOC-CRM adapted characteristics. The RSW has two main objectives. The first was to collect the attributes of any main entity within the CIDOC-CRM model by *reading* biased walks through the knowledge graph, starting with the main entities to be investigated. The second is to explicitly manifest the implicit relations between main objects within the CIDOC-CRM knowledge graph.

To accomplish the first objective of the RSW, we proceeded to initiate n walks from each of the main entities within the CIDOC-CRM knowledge graph. Given the nature of the CIDOC-CRM syntax, and the fact that a lot of relevant information is stored in the leaf nodes of relatively deep branches of the graph, we added a bias to these walks, inspired by *node2vec* [33], to favor deeper walks within the knowledge graph in search of richer information-carrying branches. In doing so, we constructed sentences in the following fashion: *head, relation, tail, relation, tail, relation, tail*.

The second objective of the RSW is accomplished by creating *relative* sentences, which make explicit the intrinsic connections between main entities. A schematic representation of such a relative walk is shown in Figure 3. Following this logic, we evaluate the importance of each entity at every step of the walk by calculating an adjustable importance metric based on an entity node centrality (see Equation (3)). If the node importance is higher than a calculated threshold θ, RSW considers the incoming relation to the current entity when evaluating its next hop destination, and can possibly make a jump, signified by the dashed red line in Figure 3, to retrace the hops within the properties of a second main entity. In this logic, the sentence generated from Figure 3 can be read as follows: *Entity A is related to entity B, which is related to entity C; entity C is a part of entity Y, which is connected to entity Z.* This approach allows us to explicitly express the implicit relations between the different main entities within the generated sentences.

RSW Algorithm

Here, we formalize the RSW approach, and present a short pseudo-code to better explain its logic. As explained above, each RSW starts from a main entity. These main entities can be chosen by the user depending on their objective, so long as the entity in question possesses at least one outgoing relation. However, given the nature of the data stored in the CIDOC-CRM knowledge graphs, such main entities tend to be objects, documents, or works of historic and cultural significance. Within the *Sphaera* dataset, the main entities can be considered to be the 359 book editions.

We considered the CIDOC-CRM knowledge graph as $G = (V, E)$, where G is the entire knowledge graph, V represents the set of nodes in this graph, which in this case is represented by all entities, and E represents the set of directed edges—in this case, relations connecting the entities within G.

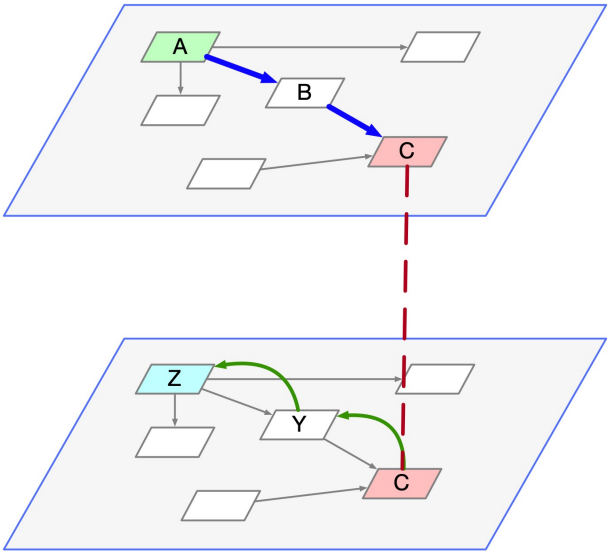

Figure 3. Schematic representation of relative sentence walks. Entities A and Z are two different main entities, along with their connecting entities in gray. Both main entities share an intrinsic connection to entity C, which is not expressed in the CIDOC-CRM logic (represented here by a red dashed line). The relative sentence walk can then reverse walk to connect entities A and Z, through C, shown here in green lines.

For each of the n random walks W_n, we start from one of the main entities, denoted here as e_0, with e_i denoting the i-th node of the walk. At each node, the transition probability can be described by Equation (1) below:

$$P(e_i|e_{i-1}) = \begin{cases} \Pi \; if \; (e_i, e_{i-1}) \in E \\ 0 \end{cases} \tag{1}$$

where Π is the normalized transfer probability between e_i and e_{i-1}.

We can control this walk by introducing a bias term β, which helps direct the transitions between entities. This bias term is constructed to follow the path of *abundant information* within the CIDOC-CRM knowledge graph. This is based on the premise that, within CIDOC-CRM, entities with a large number of outgoing relations are those that contain a larger number of relevant pieces of information, as they can lead to numerous different types of property attributes, and thus numerous types of different sentence structures. In this way, we construct β as follows in Equation (2):

$$\beta_{(i,j)} = \frac{d_{out(j)}}{\sum_j^n d_{out}} \tag{2}$$

where $\beta_{i,j}$ represents the bias term applied to the transition probability between nodes i and j, $d_{out(j)}$ represents the out-degree of node j, and the sum of d_{out} represents the local total of out-degrees from all the entities connected to node i. In this fashion, β can be thought of as a bias weight based on a local normalized out-degree which attracts the RSW towards entities with more information.

The final step of the RSW algorithm is the *relative* sentence generation. Thus, in addition to what is described above, at each step of the walk, the current node entity is investigated by examining its importance within the knowledge graph (Figure 4). The importance here is gauged again based on the entity's in and out degree. However, it is

worth mentioning that there exists several graph neural network approaches that estimate a node's importance within knowledge graphs such as GENI [34]; here, we only rely on node relation information from its direct neighbors to generate un-learned importance estimates.

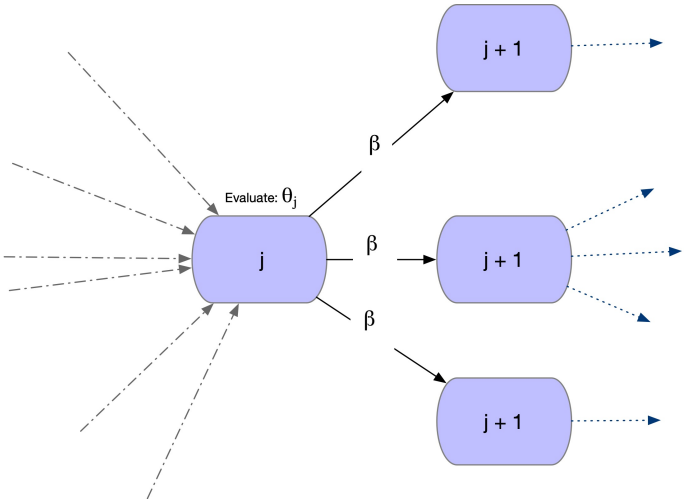

Figure 4. Node walk evaluation at node *j*.

To estimate node importance, we use node centrality, estimated by each node's in-degree, d_{in}, as shown in Equation (3) below:

$$\theta_j = \beta_j \times \log\left(d_{in,j} + \eta\right) \tag{3}$$

where η is a small positive constant and β is the weight used in the previous hop towards node j calculated in Equation (2). The use of the out-degree value based on β in Equation (3) is based on the premise that, within the CIDOC-CRM syntax, entities connected to numerous attributes are able to portray a richer semantic meaning than those that are connected to a low number of attributes (i.e., leaf nodes). It is thus beneficial to reduce the importance of leaf nodes, while highlighting the importance of node entities with the ability to carry more information. This is best demonstrated within the *Sphaera* data structure, where leaf nodes such as E56:Language can be shared by a large number of books, while entities with a high number of attributes, such a Sphaera:Part, can portray more semantically relevant information, and thus two books sharing the same part tend to have more in common than two books only sharing the same language (as seen in Figure 1).

Nodes with importance scores above a predefined threshold are thus eligible to be transition nodes, and a gateway towards the construction of a *relative sentence*. In such cases where importance scores are above the threshold, the algorithm considers outgoing edges, as well as incoming edges from other main entities, when evaluating its next move within the knowledge graph. If an incoming edge is selected, then the walk retraces its steps within the graph of the second main entity, thus connecting the two main entities by constructing meaningful sentences through the knowledge graph (see Figure 3).

The generated sentences can then take two main forms following the diagram shown in Figure 1:

- **Regular sentence**: *Book:A*—Carries—a Publication Expression—incorporates a Self Contained Expression—has Language—*Latin*.
- **Relative Sentence**: *Book:A*—Carries a—Publication Expression—incorporates—a Self Contained Expression—has component—*Part:C*—which is component—of a self-contained expression—incorporated in—a Public Expression—carried by—*Book:B*.

5.2. Document Representation Using DOC2VEC

Using the RSW algorithm, we generated n walks starting from each main entity, which created an n sentence representative document for each main entity. Inspired by

word2vec [35] and *doc2vec* [36], we generated document embeddings using *doc2vec*. This approach was based on the *word2vec* model which predicts target words based on its context, i.e., a window of words before and after the word in question. The model's input is represented by the vector representation of context words, which are then weighed, averaged, and projected in a projection layer. Using the weights from the output layer, a score was calculated for each word, which represents the probability that the chosen word is the next one in the sequence of words [36]. This is formalized as follows: $w_1, w_2, \ldots, w_n$ represents a sequence of training words, with a context window c, which then aims to maximize the average log probability shown in Equation (4):

$$\frac{1}{T}\sum_{t=1}^{T} logp(w_t|w_{t-c}...w_{t+c}) \tag{4}$$

The prediction task is often performed using Softmax, as shown in Equation (5):

$$p(w_t|w_{t-c}, ..., w_{t+c}) = \frac{e^{y_{wt}}}{\sum_i e^{y_i}} \tag{5}$$

where y_i is the un-normalized log probability for each output word i computed in Equation (6).

$$y = b + Uh(w_{t-c}, ..., w_{t+c}; W) \tag{6}$$

where U and b are the softmax parameters, and h is constructed by the concatenation of word vectors.

The above describes the *word2vec* [35] approach, where each word in a sentence, within the predefined context window, c, contributes to the prediction of the next word of that sentence. Throughout this learning process, and as an indirect result, the model learns semantic relations between words. Based on the same premise, *doc2vec* introduced a document or paragraph vector which also contributes to the learning process of documents in the same fashion as words contribute to the learning process in *word2vec* [36]. This can be thought of as another word in the vocabulary with slight differences. For example, the document vector is not shared among different documents, but is unique to each document, while word vectors are shared among different documents, such as the vector representing E56:Language *Latin*, which is the same in all the different documents where it appears. The model framework is shown in Figure 5. The model is trained using stochastic gradient descent where the gradient is obtained through back-propagation. Once trained, the document vector can be considered as a document feature vector [36], which in our case is a feature vector representative of a main CIDOC-CRM entity. The CIDOC2VEC algorithm is shown in Algorithms 1 and 2.

Algorithm 1 CIDOC2VEC.

Require: $KG(V, E) \leftarrow$ CIDOC-CRM knowledge graph, $L \leftarrow$ Unique Main Entities, $N \leftarrow$ Number of Walks per Entity, $\phi \leftarrow$ Importance Threshold, $\delta \leftarrow$ Maximum Walk Depth.
Initialize empty *Doc*
Initialize empty *DocList*
for l in L **do**
  while $n \leq N$ **do**
    Sentence = $RSW(KG, l, \phi, \delta)$
    Append *Sentence* to *Doc*
  end while
  Append *Doc* to *DocList*
end for
for *Doc* in *DocList* **do**
  Embedding = doc2vec(*Doc*)
end for

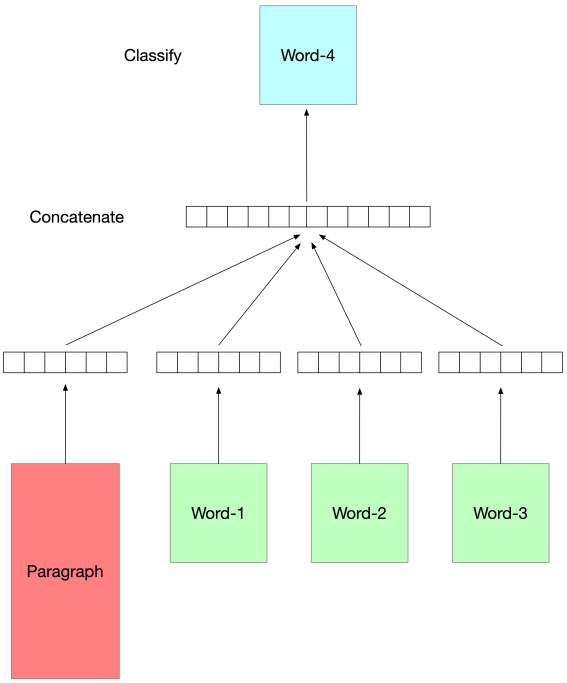

Figure 5. *doc2vec* algorithm, modified after [36].

Algorithm 2 Relative Sentence Walk, RSW.

Require: $KG(V, E) \leftarrow$ CIDOC-CRM knowledge graph, $\phi \leftarrow$ Importance Threshold, $\delta \leftarrow$ Maximum Walk Depth.
Initialize empty *Sentence*
Initialize hop counter $h = 0$
while $h \leq \delta$ **do**
  if $h \neq 0$ **then**
    Calculate θ
    if $\theta \leq \delta$ **then**
      Calculate β considering only out-edges
    else
      Calculate β considering reverse walk, in-edges.
    end if
  end if
  Evaluate hop probability and execute hop.
  Append current *Entity* to *Sentence*
  $h = h + 1$
end while
return Sentence

6. Results

The CIDOC2VEC algorithm proposed in the above section was applied to the *Sphaera* dataset both to generate an edition similarity recommendation for the *Sphaera* database (db.sphaera.mpiwg-berlin.mpg.de, accessed on 1 December 2021), and investigating the underlying patterns within the stored data. As discussed in Section 4, the database hosts a total of 359 book editions, which were decomposed and stored in a knowledge graph modeled according to the CIDOC-CRM principles (as seen in Figure 1).

We must, however, address the issue of the absence of a clear similarity benchmark when it comes to humanities data. This is of course not due to the lack of humanities datasets, but to the impossibility of creating such a clear benchmark inherent to the heterogeneous nature of humanities data. We can of course measure the similarity between edition embeddings within the embedding space. However, how do we judge whether

two editions are alike? How do we decide whether two editions are dissimilar? To what degree are they similar or dissimilar? These measures remain abstract when dealing with a humanities dataset, whether we are dealing with the *Sphaera* dataset or dealing with data whose notion of similarity is not inherent, or not simply specified. In this paper, we assess the usefulness of the CIDOC2VEC algorithm by providing a short historical overview of several groups of book editions, as well as text-parts, that historically appear to convey similar information.

6.1. Sphaera Editions

The output of the CIDOC2VEC algorithm for this test result is a 32-dimensional vector generated from 500 sentence-long documents per main entity, whose T-distributed stochastic neighbor embedding (t-SNE) [37] representation is shown in Figure 6a. Each edition is represented by its *Sphaera* Book ID, which can be used to search the publicly available database. The complete *Sphaera* corpus embedded representation, shown in Figure 6a, is clearly divided into several large clusters and within each, one can see several smaller clusters. This allows for a top–down analysis of the different components of the dataset.

We first investigated the effect of the book types on the embeddings of each entity. This is shown in Figure 6b, where each color represents a single Book type from the *Sphaera* dataset. It is clearly visible that some, but not all, of the groups can be represented by clear clusters in the t-SNE visualization. The most apparent of these form the cluster in purple representing the adaption treatises that were strongly influenced by the *Tractatus de Sphaera*, as well as the cluster of original text shown in blue. The most dispersed cluster is represented by the red color, which represents the annotated original texts and compilations. On this higher, meta-level, we can see that the embeddings clearly follow the structural format of the books, but are still highly influenced by its content. For example, while original texts contain very similar content, namely the original content of the *Tractatus de sphaera*, the annotated original texts and compilations of texts vary highly in their textual content as they include compilations of texts curated by numerous different publishers; they also include a highly variable number of authors and contributors. This such high variability results into a highly dispersed cluster of this type, which leads to the inherent data variability within this type of book.

At a more specific level, we investigated the distribution of books which contained parts written by two influential authors. The first was the German mathematician and astronomer Christopher Clavius (1538–1612) (hdl.handle.net/21.11103/sphaera.100732, accessed on 1 December 2021) who authored several influential text editions throughout their lifetime. Namely, his commentary on Euclid's *Elements* which became the standard geometry textbook in the 17th century and earned him the title of "Euclid of their time" [38], and his commentary on the *Tractatus de Sphaera*, known as *In Sphaeram Ioannis de Sacro Bosco, Commentarius* (hdl.handle.net/21.11103/sphaera.100365, accessed on 1 December 2021) which was reprinted numerous times between the second half of the 16th and the first half of the 17th centuries [39]. The second is the German Lutheran reformer and theologian Philipp Melanchthon (1479–1560) (hdl.handle.net/21.11103/sphaera.101002, accessed on 1 December 2021) who played an important role in reforming the content and structure of German learning in the 16th century, and emphasized the role of mathematics and astronomy within the university curriculum [40]. His works included numerous editions on a wide range of topics such as physics, astronomy, history, ethics, and theology, and his commentary (hdl.handle.net/21.11103/sphaera.100138, accessed on 1 December 2021) on the *Sphaera de Sacrobosco* was reprinted throughout the 16th and first half of the 17th centuries [41]. These two authors contributed to a substantial number of editions within the *Sphaera* corpus. Editions with parts authored by these two individuals are shown in green and red in Figure 6c, and their embeddings are clearly clustered according to the coherent content, as well as the types of books in which they often feature. Editions by Christopher Clavius, with the exception of three texts, appear in a single cluster, highlight-

ing the homogeneity of the editions in which his work features. The three editions that do not conform to the cluster in question appear to be a single book containing a Spanish translation of the original Latin work (hdl.handle.net/21.11103/sphaera.100555, accessed on 1 December 2021), while the other two editions contain only some passages of Clavius' text, published together with further passages from a number of different commentaries (hdl.handle.net/21.11103/sphaera.100327, accessed on 1 December 2021). The editions containing texts by Philipp Melanchthon represent a more coherent cluster that stretches across different book types, namely, books of annotated originals text and compilations of texts, along with books solely containing compilations of texts. However, due to the similarity of the content, as well as the closeness of the structure of the two types of books in which Philipp Melanchthon's texts feature, these are embedded in the close proximity, and thus form a coherent, similar, body of work.

Finally, we investigated the phenomenon of book reprinting in consecutive years which was strictly related to the production and commercial praxis of early modern printers and publishers. Due to the production-inherent necessity of producing oversized print-runs, it was usual for book producers to sell similar exemplars in consecutive years, each time branded as a new edition, by solely printing a new title page (for production procedures, see [42]; for marketing practices, see [43]). While this book reprinting phenomenon is not explicitly expressed in the *Sphaera* knowledge graph, its effects are clearly recognizable in the embedding space through very dense clusters, resulting from the fact that these books often contained the same content and are sometimes reproduced by the same individuals over a relatively long period of time, with only minute changes. Two of these reprint families are marked by dashed boxes in Figure 6a. The first set of re-print families is shown in the red dashed box, and consists of a total of nine reprints of the *Elementae doctorinae de circulis coelestibus, et primo motu* (hdl.handle.net/21.11103/sphaera.100251, accessed on 1 December 2021) authored by Kaspar Peucer (1525–1602) (hdl.handle.net/21.11103/sphaera.100966, accessed on 1 December 2021), which was reprinted between 1551 and 1601. The second re-print family is represented in the blue box and represents a total of 11 reprints of Thomas Blebel's (1539–1596) (hdl.handle.net/21.11103/sphaera.100342, accessed on 1 December 2021) *De Sphaera et primis astronomiae rudimentis*. (hdl.handle.net/21.11103/sphaera.101095, accessed on 1 December 2021) between 1576 and 1629. While it is apparent that many more of such reprint families appear in the dataset, it was not the aim of this paper to discuss the mechanism of early modern period book production.

6.2. Sphaera Text-Parts

The results above show the capabilities of CIDOC2VEC in both generating meaningful embeddings from complete CIDOC-CRM knowledge graphs, as well as explicitly expressing implicit information such as those expressed by the reprinting families. However, to better investigate CIDOC2VEC's agility, we explore the same *Sphaera* dataset by using parts, instead of books, as main entities so that every generated walk starts from a part rather than a book entity. Additionally, we omit certain relations from our knowledge graph, namely the part type and part identifier relations. This way, we explore an incomplete knowledge graph, where the main entity data type is unknown on the one hand, while on the other, we explore the performance of CIDOC2VEC on a peripheral main entity (see Figure 1), as opposed to the central location occupied by the book entity. Passing the now incomplete knowledge graph to the CIDOC2VEC algorithm, we obtained a 32-dimensional vector, generated from 500 walks, for each of the 450 text-part main entities, whose t-SNE plot is shown in Figure 7, where each entity is identified by its *Sphaera* Part ID and validated using our hold-out type relations represented by different colors.

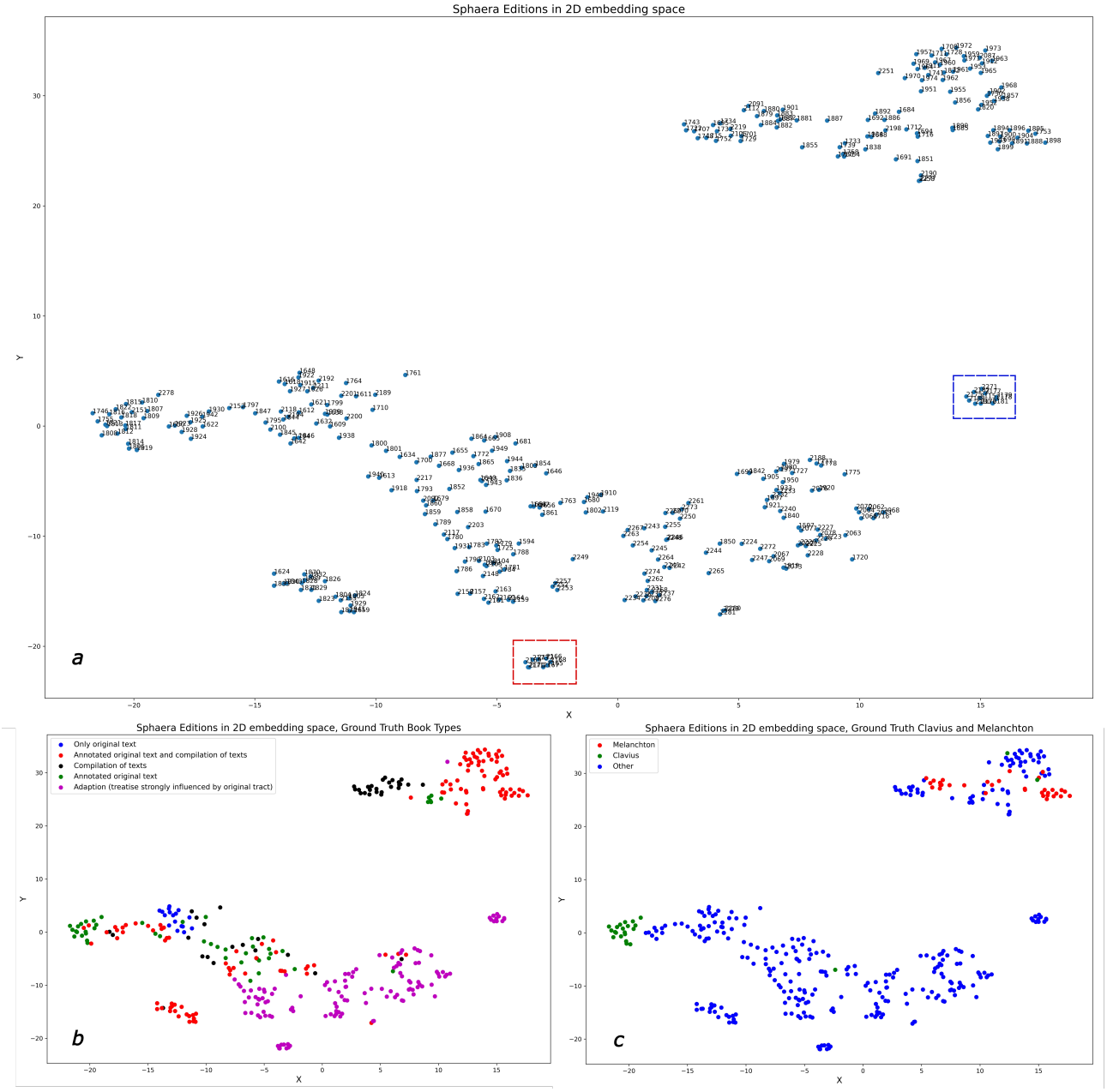

Figure 6. (**a**): t-SNE representation of all 359 book edition entities in the *Sphaera* dataset; (**b**): ground truth data-based book types; (**c**): ground truth data based on books containing texts by specific authors.

The t-SNE plot clearly shows several clusters, mostly coherent and formed from distinct part types, with the exception of Paratext:Other entities which show a level of dispersion as a result of its inherent incoherence. The largest, and most important, cluster is one represented by content class which forms the bulk of the text within the *Sphaera* corpus. We notice that several entities of different classes are located on the periphery of the content cluster. Upon closer investigation, we see that these entities are in fact some of the most frequently published paratexts in the corpus, and often written by influential authors from the *Sphaera* corpus. Consequentially, these paratext parts are continuously associated with content parts due to their higher publication frequency, as well as through their author association. Such examples include a poem by Philipp Melanchton titled *de triplici ortu*(hdl.handle.net/21.11103/sphaera.100849, accessed on 1 December 2021) , which was printed a total of 39 times in the 16th and at the beginning of the 17th centuries, as well as a letter to the reader, simply titled *ad lectorem* (hdl.handle.net/21.11103/sphaera.100359,

accessed on 1 December 2021), by Christopher Clavius, published 20 times in the second half of the 16th and the first half of the 17th century.

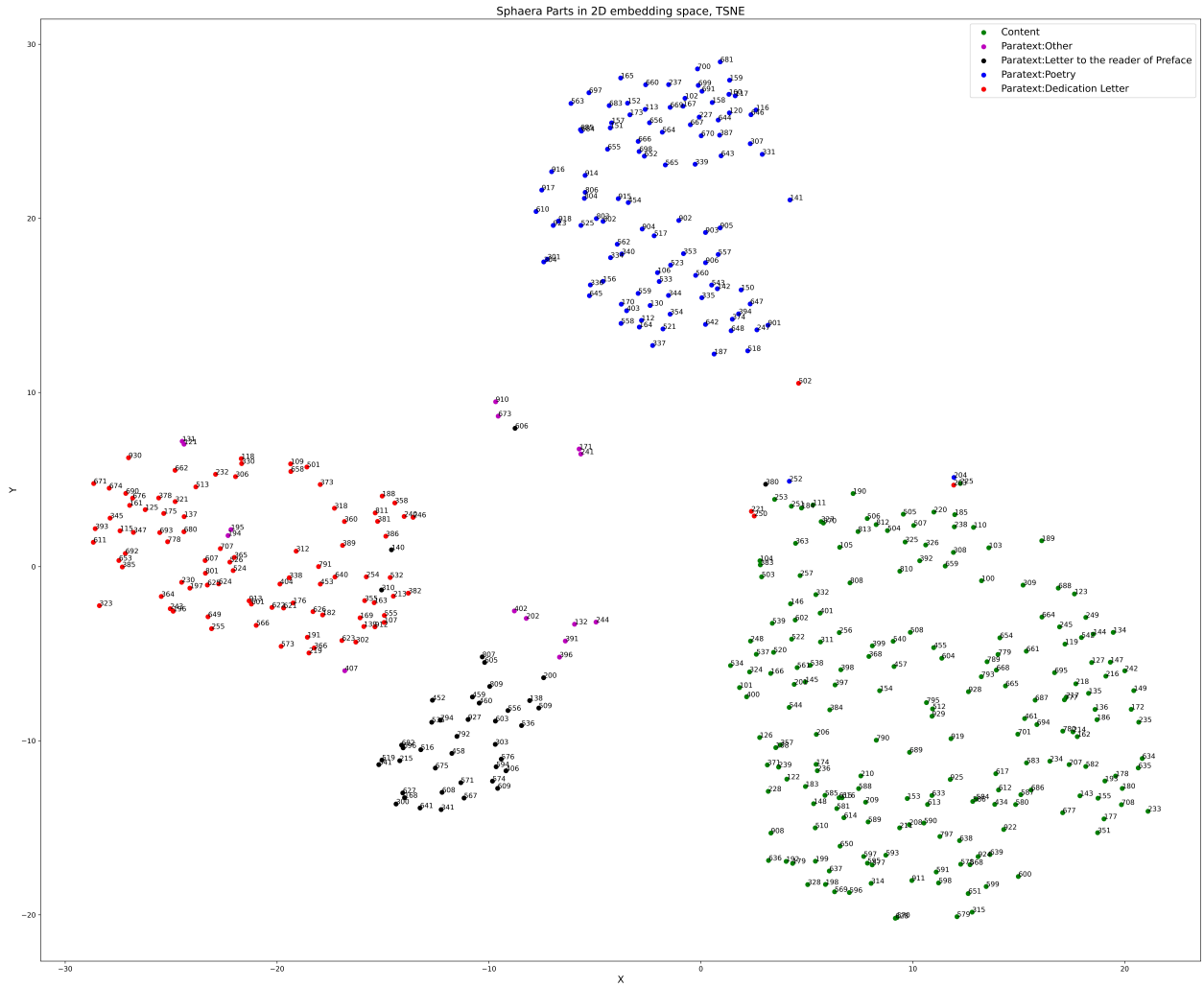

Figure 7. The t-SNE representation of the 450 part entities generated from an incomplete *Sphaera* knowledge graph.

Additionally, we investigated the cluster representing the Paratext:Dedication letter class, shown in red in Figure 7. While this cluster is relatively clean, the presence of out-of-class entities at its center, namely those representing the class Paratext:Other, is apparent. However, we notice that those Paratext:Other entities are also in fact letters, but differ semantically from dedication letters. This semantic difference lies in the functionality of the letter in question; while all the letters within the Paratext:Dedication letter class serve to dedicate the content of the edition to a certain individual, the few letters within the Paratext:Other class often show the correspondence between authors and the responsible authorities, in which the former is asking for the permission to publish the work, known as the imprimatur, from the latter. The low number of such types of letters in the *Sphaera* dataset is due to the fact that, while it was imperative to obtain permission to publish a certain work, it was not imperative to print the correspondence. Consequently, the presence of these out-of-class letters within the Paratext:Dedication letter cluster clearly indicates that these entities share a lot of their semantic and structural data, but are still differentiated through the historical interpretation of their texts, which are themselves not part of the *Sphaera* Corpus. In this regard, the results obtained from passing an incomplete CIDOC-CRM knowledge graph to CIDOC2VEC shows the ability of our algorithm to express and materialize inherent and un-instantiated relations through its embeddings,

which can guide domain experts towards a better understanding of their CIDOC-CRM modeled data.

6.3. Model Stability

Having demonstrated the ability of CIDOC2VEC to extract meaningful embeddings from complete and incomplete CIDOC-CRM knowledge graphs, in this section, we address the effects of the randomness present in the algorithm due to the biased random walks of the RSW. We accomplish this by running the CIDOC2VEC algorithm 10 times on the complete *Sphaera* knowledge graph and evaluating the stability of the local neighborhoods, which was calculated as the ratio of the intersection between the top-10 neighborhood of every main entity over multiple runs. We demonstrated this in Figure 8, which shows the neighborhood stability measure of a subset of *Sphaera* editions, which includes those discussed in Section 6.1, as well as their overall stability within the *Sphaera* corpus.

It is apparent that, while some of the entities present a consistently high stability score, others show a relatively unstable neighborhood over the course of the 10 CIDOC2VEC runs. This is, however, to be expected, and can be explained by the nature of the editions in question. While, for example, the editions authored by Christoph Clavius and Philipp Melanchthon, as well as the reprinting families by Kaspar Peucer and Thomas Blebel show a consistently high stability score, mainly due to the fact that they are part of highly homogenized clusters, as shown in Figure 6 and discussed in Section 6.1, other editions cannot be attributed to any specific families or clusters within the *Sphaera* dataset. This is especially the case for editions written in under-represented languages, such as Spanish or Portuguese, which are represented by a total of 10 and 3 editions, respectively. This is clearly visible in Figure 8, where the edition authored by Francisco Faleiro (1490 – 1550) (hdl.handle.net/21.11103/sphaera.101184, accessed on 1 December 2021), written in Spanish and titled *Tratado del Esphera y del arte del marear* (hdl.handle.net/21.11103/sphaera.101182, accessed on 1 December 2021), as well the edition authored by an anonymous author (hdl.handle.net/21.11103/sphaera.101356, accessed on 1 December 2021) writen in Portuguese and titled *Regimento do estrolabio e do quadrante* (hdl.handle.net/21.11103/sphaera.100916, accessed on 1 December 2021) show low stability, arising from the fact that they show relatively little similarity to the rest of the corpus. However, it is clear from the stability of the corpus average, shown by a dotted black line in Figure 8 and hovering around 0.78, that such unstable cases are a minority and do not represent a general trend in the *Sphaera* dataset. Instead, the general stability of the neighborhoods indicate that with every run of CIDOC2VEC, we are able to extract the same semantic relations between the main entities, despite the randomization injected by the RSW.

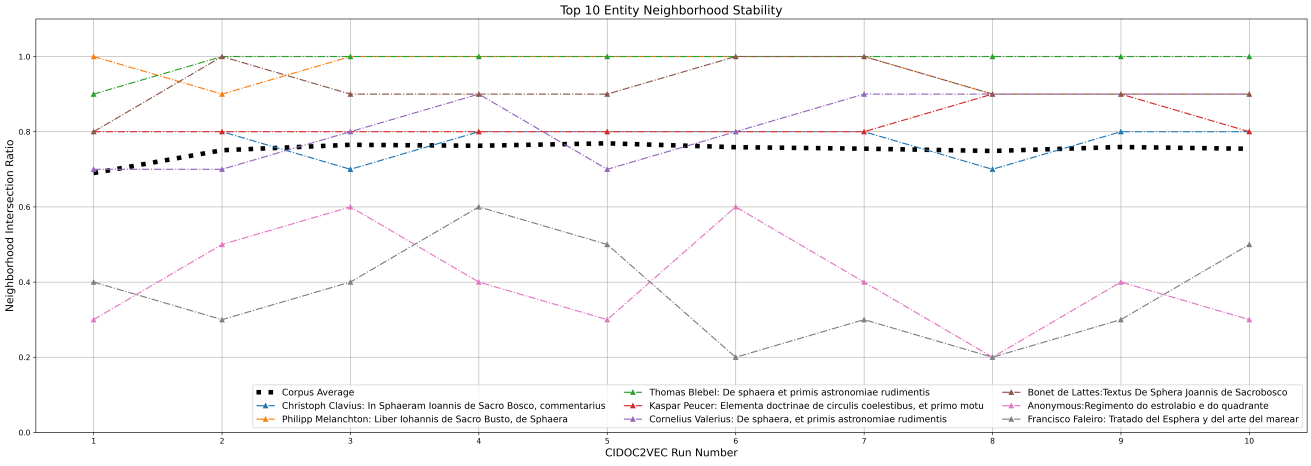

Figure 8. Main entity neighborhood stability over 10 CIDOC2VEC runs.

7. Conclusions

Through the demonstration of the CIDOC2VEC using the *Sphaera* dataset, and the historical confirmation of its results, we have showed that this approach is useful to generate meaningful, persistent, and representative embeddings of humanities datasets stored in CIDOC-CRM knowledge graph structures. These embeddings lead to accurate similarity suggestions between CIDOC-CRM entities based on a representative set of information that has been collected using the relative sentence walk through the CIDOC-CRM knowledge graph. Thus, materializing the intrinsic connections between otherwise un-connected entities helps one generate closer embeddings of related objects. Such similarity suggestions can help guide users of humanities databases to discover similar items based on the underlying data model. Additionally, we showed that such representative embeddings help users discover hidden patterns within the dataset that would be difficult to conceive by only looking at the individual entities. While we acknowledge that it is not possible to provide an accurate metric of how two entities are similar within the humanities, this remains one of the major challenges of applying data-centric approaches to humanistic problems. Our aim here was to present a tool that can help researchers working with humanities data to look deeper into their own datasets and extract more information from them in order to reach more informed conclusions.

Author Contributions: Conceptualization, H.E.-H. and M.V.; methodology, H.E.-H.; software, H.E.-H.; validation, H.E.-H. and M.V.; data curation, H.E.-H.; writing, H.E.-H. and M.V.; visualization, H.E.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the German Ministry for Education and Research as BIFOLD—Berlin Institute for the Foundations of Learning and Data (ref. 01IS18037A), the Max Planck Institute for the History of Science, and the Max Planck Society.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The *Sphaera* public database can be accessed here: http://db.sphaera.mpiwg-berlin.mpg.de (accessed on 1 December 2021). The cidoc2vec code can be accessed here: https://gitlab.gwdg.de/MPIWG/Department-I/sphaera/cidoc2vec (accessed on 1 December 2021) and https://github.com/hassanhajj910/cidoc2vec (accessed on 1 December 2021).

Acknowledgments: The authors would like to thank Nathaniel LaCelle-Peterson and Kim Pham for proofreading the text.

Conflicts of Interest: The authors declare no conflict of interest. The founders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Wang, X.; He, X.; Cao, Y.; Liu, M.; Chua, T.S. KGAT: Knowledge Graph Attention Network for Recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 950–958.
2. Christmann, P.; Saha Roy, R.; Abujabal, A.; Singh, J.; Weikum, G. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 729–738.
3. Kraütli, F.; Valleriani, M. CorpusTracer: A CIDOC Database for Tracing Knowlege Networks. *Digit. Scholarsh. Humanit.* **2018**, *33*, 336–346. [CrossRef]
4. Görz, G.; Seidl, C.; Thiering, M. Linked Biondo: Modelling Geographical Features in Renaissance Texts and Maps. *e-Perimetron Int. Web J. Sci. Technol. Affin. Hist. Cartogr. Maps* **2021**, *16*, 78–93.
5. Koho, M.; Ikkala, E.; Leskinen, P.; Tamper, M.; Tuominen, J.; Hyvönen, E. WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. *Semantic Web* **2021**, *12*, 265–278. [CrossRef]

6. Sinikallio, L.; Drobac, S.; Tamper, M.; Leal, R.; Koho, M.; Tuominen, J.; La Mela, M.; Hyvönen, E. Plenary Debates of the Parliament of Finland as Linked Open Data and in Parla-CLARIN Markup. In Proceedings of the 3rd Conference on Language, Data and Knowledge (LDK 2021), Zaragoza, Spain, 1–3 September 2021; Gromann, D., Sérasset, G., Declerck, T., McCrae, J.P., Gracia, J., Bosque-Gil, J., Bobillo, F., Heinisch, B., Eds.; Schloss Dagstuhl—Leibniz-Zentrum für Informatik: Dagstuhl, Germany, 2021; Volume 93, pp. 8:1–8:17.
7. Mäkëla, E.; Törnros, J.; Lindquist, T.; Hyvönen, E. WW1LOD: An application of CIDOC-CRM to World War 1 linked data. *Int. J. Digit. Libr.* **2015**, *18*, 333–342. [CrossRef]
8. Felicetti, A.; Murano, F. Scripta Manent: A CIDOC CRM Semiotic Reading of Ancient Texts. *Int. J. Digit. Libr.* **2017**, *18*, 263–270. [CrossRef]
9. Haslhofer, B.; Isaac, A.; Simon, R. Knowledge Graphs in the Libraries and Digital Humanities Domain. In *Encyclopedia of Big Data Technologies*; Sakr, S., Zamaya, A., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 1–8.
10. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*; Aberer, K., Choi, K.S., Noy, N., Allemang, D., Lee, K.I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., et al., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
11. Vrandečić, D.; Krötzsch, M. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]
12. Bizer, C.; Heath, T.; Berners-Lee, T. Linked Data—The Story So Far. *Int. J. Semantic Web Inf. Syst.* **2009**, *5*, 1–22. [CrossRef]
13. Bekiari, C.; Bruseke, G.; Doerr, M.; Ore, C.E.; Stead, S.; Velios, A. Definition of the CIDOC Conceptual Reference Model v7.1.1. 2021. Available online: https://cidoc-crm.org/sites/default/files/cidoc_crm_v.7.1.1_0.pdf (accessed on 1 December 2021).
14. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2013; Volume 26.
15. Wang, Z.; Zhang, J.; Feng, J.; Chen, Z. Knowledge Graph Embedding by Translating on Hyperplanes. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014; pp. 1112–1119.
16. Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 2181–2187.
17. Ji, G.; He, S.; Xu, L.; Liu, K.; Zhao, J. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Association for Computational Linguistics: Beijing, China, 2015; pp. 687–696.
18. Fan, M.; Zhou, Q.; Chang, E.; Zheng, T.F. Transition-based Knowledge Graph Embedding with Relational Mapping Properties. In Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing, Phuket, Thailand, 12–14 December 2014; pp. 328–337.
19. Xiao, H.; Huang, M.; Hao, Y.; Zhu, X. TransA: An Adaptive Approach for Knowledge Graph Embedding. *arXiv* **2015**, arXiv:abs/1509.05490.
20. Dain, Y.; Wang, S.; Xiong, N.; Guo, W. A Survey of Knowledge Graph Embedding: Approaches, Applications, and Benchmarks. *Electronics* **2020**, *9*, 750.
21. Nickel, M.; Tresp, V.; Kriegel, H.P. A Three-Way Model for Collective Learning on Multi-Relational Data. In Proceedings of the 28th International Conference on International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; Omnipress: Madison, WI, USA, 2011; pp. 809–816.
22. Yang, B.; Yih, W.; He, X.; Gao, J.; Deng, L. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
23. Bordes, A.; Glorot, X.; Weston, J.; Bengio, Y. A semantic matching energy function for learning with multi-relational data. *Mach. Learn.* **2014**, *94*, 233–259. [CrossRef]
24. Nguyen, D.Q.; Nguyen, T.D.; Nguyen, D.Q.; Phung, D. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 327–333.
25. Socher, R.; Chen, D.; Manning, C.D.; Ng, A. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In *Advances in Neural Information Processing Systems*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2013; Volume 26.
26. Meghini, C.; Doerr, M. A first-order logic expression of the CIDOC conceptual reference model. *Int. J. Metadata Semant. Ontol.* **2018**, *13*, 131–149. [CrossRef]
27. Valleriani, M.; Kraütli, F.; Zamani, M.; Tejedor, A.; Sander, C.; Vogl, M.; Bertram, S.; Funke, G.; Kantz, H. The Emergence of Epistemic Communities in the *Sphaera* Corpus: Mechanisms of Knowledge Evolution. *J. Hist. Netw. Res.* **2019**, *3*, 50–91.
28. Bekiari, C.; Doerr, M.; Boeuf, P.L.; Riva, P. Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism. 2015. Available online: https://repository.ifla.org/handle/123456789/659 (accessed on 23 October 2021).
29. Zamani, M.; Tejedor, A.; Vogl, M.; Kraütli, F.; Valleriani, M.; Kantz, H. Evolution and transformation of early modern cosmological knowledge: A network study. *Sci. Rep.* **2020**, *10*, 19822. [CrossRef] [PubMed]
30. Toutanova, K.; Chen, D.; Pantel, P.; Poon, H.; Choudhury, P.; Gamon, M. Representing Text for Joint Embedding of Text and Knowledge Bases. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 1499–1509.

31. Liang, S.; Kurt Stockinger, T.M.; Anisimova, M.; Gil, M. Querying Knowledge Graphs in Natural Language. *J. Big Data* **2021**, *8*, 3. [CrossRef] [PubMed]
32. Agarwal, O.; Ge, H.; Shakeri, S.; Al-Rfou, R. Large Scale Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training. *arXiv* **2020**, arXiv:abs/2010.12688.
33. Grover, A.; Leskovec, J. Node2vec: Scalable Feature Learning for Networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 855–864.
34. Park, N.; Kan, A.; Dong, X.L.; Zhao, T.; Faloutsos, C. Estimating Node Importance in Knowledge Graphs Using Graph Neural Networks. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 596–606.
35. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
36. Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 21–26 June 2014; Xing, E.P., Jebara, T., Eds.; PMLR: Bejing, China, 2014; Volume 32, pp. 1188–1196.
37. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
38. Lattis, J. *Between Copernicus and Galileo: Christoph Clavius and the Collapse of the Ptolemaic Cosmology*; The University of Chicago Press: Chicago, IL, USA, 1994.
39. Sigismondi, C. Christopher Clavius astronomer and mathematician. *Il Nuovo C.* **2012**, *36*, 231–236.
40. Brosseder, C. *Im Bann der Sterne: Caspar Peucer, Philipp Melanchthon und andere Wittenberger Astrologen*; Akademie Verlag: Berlin, Germany, 2004.
41. Westman, R.S. The Melanchthon Circle, Rheticus, and the Wittenberg Interpretation of the Copernican Theory. *Isis* **1975**, *66*, 165–193. [CrossRef]
42. Werner, S. *Studying Early Printed Books, 1450–1800: A Practical Guide*; Wiley Blackwell: Hoboken, NJ, USA, 2019.
43. Maclean, I. *Episodes in the Life of the Early Modern Learned Book*; Brill: Leiden, The Netherlands, 2020.