Datenbank Spektrum
https://doi.org/10.1007/s13222-022-00414-1

SCHWERPUNKTBEITRAG

An Ever-Expanding Humanities Knowledge Graph: The *Sphaera* Corpus at the Intersection of Humanities, Data Management, and Machine Learning

Hassan El-Hajj[1,2] · Maryam Zamani[1,2,3] · Jochen Büttner[1,2] · Julius Martinetz[1,2,4] · Oliver Eberle[2,4] · Noga Shlomi[1,5] · Anna Siebold[1,6] · Grégoire Montavon[2,4] · Klaus-Robert Müller[2,4,7,8] · Holger Kantz[3] · Matteo Valleriani[1,2,4,5]

Received: 15 February 2022 / Accepted: 22 April 2022

Abstract

The Sphere project stands at the intersection of the humanities and information sciences. The project aims to better understand the evolution of knowledge in the early modern period by studying a collection of 359 textbook editions published between 1472 and 1650 which were used to teach geocentric cosmology and astronomy at European universities. The relatively large size of the corpus at hand presents a challenge for traditional historical approaches, but provides a great opportunity to explore such a large collection of historical data using computational approaches. In this paper, we present a review of the different computational approaches, used in this project over the period of the last three years, that led to a better understanding of the dynamics of knowledge transfer and transformation in the early modern period.

Keywords Digital Humanities · Early Modern Period · Machine Learning · Knowledge Evolution · Data Management · Network Analysis · Explainable Artificial Intelligence

✉ Hassan El-Hajj
hhajj@mpiwg-berlin.mpg.de

[1] Max Planck Institute for the History of Science, Berlin, Germany

[2] BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany

[3] Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

[4] Technische Universität Berlin, Berlin, Germany

[5] Tel-Aviv University, Tel-Aviv, Israel

[6] Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

[7] Department of Artificial Intelligence, Korea University, Seoul, Korea (Republic of)

[8] Max Planck Institute for Informatics, Saarbrücken, Germany

1 Introduction

The project *Sphere: Knowledge System Evolution and the Shared Scientific Identity of Europe*[1] aims to study the evolution of knowledge throughout the early modern period by investigating a total of 359 university textbook editions printed between 1472 and 1650, centered on the *Tractatus de sphaera* by Johannes de Sacrobosco (–1256). The corpus allows us to reconstruct the evolution of scientific knowledge in Europe for four consecutive centuries. Despite the relative simplicity of the treatise's content, its importance to understanding the evolution of knowledge stems from the fact that its continuous transformation and modification, through commentaries and adaptions and from the thirteenth to the seventeenth century, allow us to investigate the broader mechanisms of knowledge evolution during this period.

Each of the 359 editions is represented by a single digital copy that is considered to be a representative sample of the entire edition print-run, resulting in a corpus that contains almost 76,000 pages. While such a corpus presents a challenge to traditional historical approaches, the abundance of

[1] https://sphaera.mpiwg-berlin.mpg.de.

Published online: 16 May 2022

data created a great opportunity for the application of computational methods to understand mechanisms of the evolution, homogenization, and mathematization of scientific knowledge during the phase of the emergence of modern science throughout early modern Europe.

Merging computational and historical aspects places the Sphere project at the center of the emerging field of Digital Humanities (DH), which encompasses a wide array of computational research to better analyze the ever increasing amount of historical datasets. Such projects range from those which rely on a combination of knowledge atomization and an underlying ontology to model historical events [9, 12], while in [28], the *Republic of Letters* is studied through a multi-layered network of its books. Deep-learning approaches are used in [1, 17] with the aim of studying the visual material of historical newspapers as well as dating the writing style of medieval manuscripts. While a comprehensive review of the DH landscape is beyond the scope of this paper, the following sections discuss the numerous computational approaches used in the Sphere project.

2 *Sphaera* Knowledge Graph and Knowledge Atoms

The rigorous historical analyses that form the foundation of the Sphere project brought to the identification of five different classes of editions within the corpus, clearly differentiated by their content. The “original treatises” class represents a total of 17 editions which exclusively contain the original text of the *Tractatus de sphaera* without added contemporary commentaries. 48 other editions, classified as “annotated original treatises,” contain the original work of Johannes de Sacrobosco with additional commentaries by various authors. In “Compilation of texts,” we define a class of 43 editions which include the original *Tractatus de sphaera* along with other original treatises by various authors, while the class “compilation of text and annotated originals” contains 124 editions which include a commented or annotated *Tractatus de sphaera* along with other treatises. The final and largest class is constituted by editions defined as “adaptions”, which number 127 and display texts that are strongly influenced by the content and structure of the *Tractatus de sphaera*, but do not include the original treatise itself.

In order to better explore these editions, we perform what we refer to as multi-level edition atomization. We identify text-parts, scientific illustrations, and numerical computational astronomic tables as ‘knowledge atoms’. On the first level, each edition is atomized – or deconstructed – into text-parts, each representing a textual component that is both larger than a single paragraph but also convey a coherent body of information [26]. On a second level, we performed a content-related analysis of the text-parts in order to assess their mutual semantic relations. First distinguishing between “original texts,” “commentaries,” and “translations,” we then relate the text-parts to each other using the relationships “commentary of,” “translation of,” and “fragment of.”

The result of this analysis is stored in a knowledge graph [13], and modelled according to the CIDOC-CRM Ontology [4], which provides a useful and standardized framework for modelling and storing humanities and cultural heritage data; the framework also strives to create coherent and shareable datasets across research institutions[2]. This knowledge graph forms the basis for all further investigation of the *Sphaera* Corpus, and has expanded to be a number of times larger than its original size due to multiple consecutive historical and computational research cycles [15].

3 Multi-layer Network Analysis

In order to investigate the emergence of epistemic communities within the *Sphaera* corpus, we constructed multiple networks from the 356 *Sphaera* editions.[3] These networks are built by considering each of the 356 editions of the corpus as nodes in different layers of a multiplex network, connected by different types of edges. These editions are composed of 563 different text-parts, from which 239 are reprinted at least once [26].

The layers of the multiplex network represent different semantic categories. For example, one layer is constructed by connecting edition nodes where a text-part was reprinted without changes, while others are constructed by creating connections based on the relations “text-part translation of,” “annotation of,” or “adaptation of.” If a later edition of the corpus contains a unit of text based on a text-part of an earlier edition, a link is created between these two editions, resulting in directed weighted multiplex network. The link direction in all the layers is set from a source to a target edition, where the former is published earlier than the latter. This structure of semantic layers discussed above, as well as additional socio-economic layers constructed based on relations between edition producers and economic factors is inherent to the database [13] and defines the network of this study.

[2] *Sphaera* database access: https://db.sphaera.mpiwg-berlin.mpg.de.

[3] At the time of this study, 356 editions were included in the *Sphaera* corpus. Three further editions were added after this study to bring the total number of *Sphaera* editions to 359.

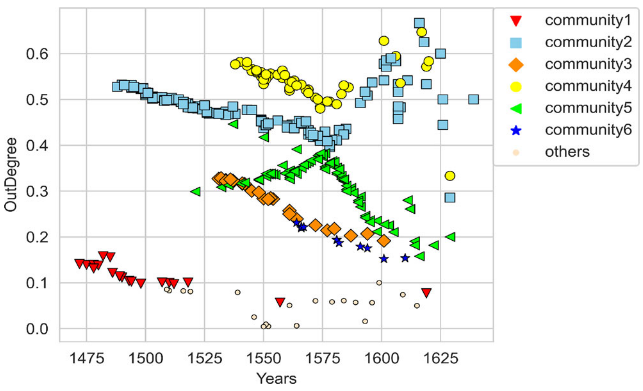

Fig. 1 Normalized out-degree as a function of publication time for the aggregated graph. Color coding shows the 6 communities obtained by means of the Louvain community detection algorithm

3.1 Network Analysis, Families of Editions, and Communities

A good insight can be gained from the aggregated graph, which is obtained through the union of all nodes, and the accumulation of the links from all layers in a single network. The influence of an edition on the subsequent development is then related to its out-degree. However, the later in time an edition is published, the lower the number of editions that will be published after it, which is evident in the lower upper-bound value of the out-degree the later one moves in time. Therefore, we normalize the out-degree of every edition by the total number of editions which are published afterwards. The result is shown in Fig. 1. The plot shows several branches which correspond to edition families with similar content [30].

A simple null-model supports our interpretation: Let us have two different versions of the *Sphaera* corpus. Let version 1 be reproduced identically with a connection to 2/3 of all later books, and version 2 with 1/3. Then every book of version 1 will have twice as many outgoing links as every book of version 2, and after normalization we will see two values of the out-degree: one family at the value 2/3, and one family at the value 1/3. Thus the family to which an edition belongs to can be identified by its normalized out-degree. In reality, the situation is more complex, since editions are not the exact reproductions of earlier ones. In addition, editions of different families contain different numbers of text-parts, which results in different out-degrees.

An alternative approach to detect edition families is network community detection. To accomplish this we used the Louvain community detection algorithm [5] and found that the branches in Fig. 1 are mostly formed by editions of a single community.

3.2 Most Influential Editions, Innovations, and Great Transmitters

We pay special attention to the innovations in science and the formation of new ways of thinking. We consider an edition as innovative if it is the beginning of a new family as shown in Fig. 1, especially if many later editions tend to refer to this particular innovative edition more often than its earlier or contemporary ones. In this case, a very likely interpretation is that some relevant content of this edition cannot be found in earlier ones, and hence an innovation has occurred. This is not a disruptive innovation however, as disruption quantifies a slightly different property. In disruption cases, an edition should not only contain an innovation, but at the same time break with past traditions. The disruptive edition itself should not have many links to the past, but have many links to the future. This is portrayed by a strong imbalance of in- and out-degrees. Our analysis revealed several particularly important editions for the evolution of knowledge. Fig. 2 shows the normalized average age of the links to the source editions (a) and the normalized average age of the links to the targets editions (b) [30]. As a result, we identified *enduring innovations* and *great transmitters*.

The *enduring innovations* are indeed visible in Fig. 1 as nodes at the beginning of a given edition family, as well as in Fig. 2 as the editions at the start of community five. These refer to the latest editions (the average age of the links to the sources is smaller than that of the reference complete graph) and their influence extends far into future (the average age of the links to the targets are higher than the reference graph) [30]. In Fig. 2, 25 editions were detected because they converge to the values of the reference graph. These editions were produced between 1549 and 1562, predominantly in Wittenberg, incorporate past knowledge through their connections to almost all the past instances, and maintain connections to later works; the content of knowledge in these editions spans almost the entire period of the corpus. We call these editions the *great transmitters*. Our network analysis is also able to detect *sleeping beauties*, i.e., text parts which where not recognized for many years after publication, and only much later became very popular and hence frequently reprinted [11]. As discussed above, editions inside each community are highly interconnected, they form different branches in the out-degree graph (Fig. 1) and have different types of properties in terms of the length of connections to past and future editions (Fig. 2). Two examples of editions in corpus as "great transmitter" and "enduring innovation" are *Iohannis de Sacro Busto Libellus de sphaera*[4] published in 1561 in

[4] https://hdl.handle.net/21.11103/sphaera.101109.

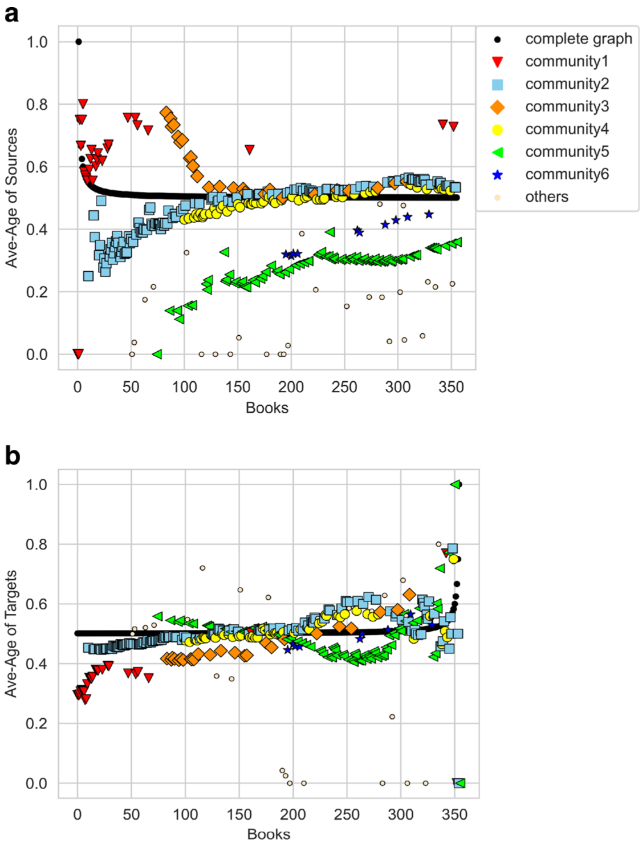

Fig. 2 **a** Normalized average age of the links referring to the source editions versus the temporal edition order in the Sphaera network. **b** Normalized average age of the links that depart from each edition towards targets versus the temporal edition order

Wittenberg and *De la Sfera del Mondo*[5] published in 1540 in Venice.[6]

4 Building and Explaining Similarity for Numerical Tables

In addition to our efforts to build ontologies for representing the heterogeneity of historical data and analyzing connectivity of the historical network, there is a further possible layer of analysis related to the content, represented here by numerical tables from the *Sphaera* corpus.

In the following, we present our work on building a large, robust, and transparent similarity model between the approximately 10,000 numerical tables of the *Sphaera* Corpus. The model we have built is a key step towards our overall goal of reconstructing the process of mathematization at the beginning of modern science. So far, comparisons between tables could only be achieved by a skilled historian, and due the difficulty of carefully and consistently inspecting the hundreds of digits composing a typical table, this approach has been strongly limited in terms of dataset size.

4.1 Machine Learning of Table Similarity

Machine learning brings the promise of scaling up the analysis of historical content to much larger corpora, in our case, the whole corpus of 10,000 numerical tables. The vast majority of ML approaches work in an end-to-end fashion [6, 23], where the prediction function is learned from the input to the output, based on output labels provided by domain experts. Due to the prohibitive cost of producing these annotations, we have proposed a bottom-up approach which dramatically reduces the labeling cost by only requiring a few annotations of some random yet representative selection of digits occurring in the table corpus [7]. These annotations serve to learn a simple single-digit 'neural OCR'. This digit recognition layer is followed by a collection of hard-coded functions that compose the detected individual digits and build some desired invariances. Specifically, our proposed neural network architecture, which we call 'bigram network', consists of three blocks: (1) A convolutional neural network trained to recognize single digits, which is slid over the whole input table to produce a digit activation map, (2) a block that multiplies adjacent locations of the activation maps to recognize pairs of adjacent digits (00-99) or 'bigrams', (3) a block that pools bigrams spatially to compute a 100-dimensional histogram representing the number of occurrences of each bigram in the table. From this histogram representation, similarity scores between tables can be obtained by computing dot products or distances in histogram space.

While the bigram representation discards a lot of (potentially relevant) information, we find that such a loss of information offers robust invariance to the page layout. At the same time, we find that the large number of digits composing each table makes up for such loss of information, and still enables a reliable similarity assessment of the different tables. The similarity model can also be used to support a t-SNE [18] low-dimensional embedding of the dataset which we show in Fig. 3, and where we can observe that pairs of numerically identical tables are embedded to nearby map locations.

4.2 Validation using Explainable AI

Explainable AI has provided machine learning with tools to go beyond common validation procedures, in particular, by revealing to the user what are the input features that contribute the most to the model prediction. This is especially important if appropriate annotation data for the

[5] https://hdl.handle.net/21.11103/sphaera.101026.

[6] For an in-depth analysis of these works and the historical meaning of these epistemic categories, see [30].

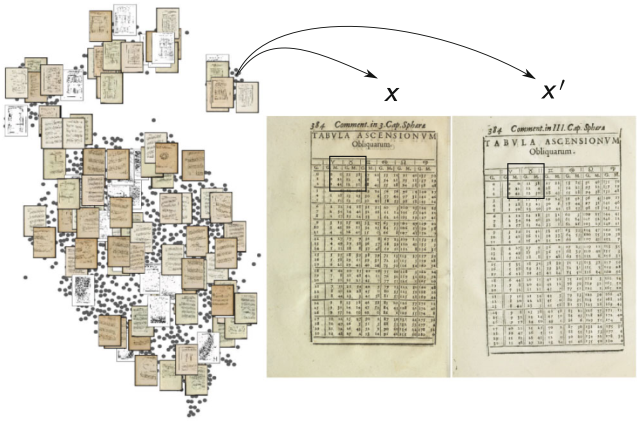

Fig. 3 T-SNE visualization of the *Sphaera* table pages, represented as histograms by the bigram network. The two highlighted pages are numerically identical

evaluation of correct model behavior is not available. For a historical analysis, we are not only interested in a well predicting model, but we also wish to verify that the conclusions drawn from a machine learning model are supported by meaningful data features and not, for example, by confounding variables. Hence, it is desirable to make the model transparent, in particular, features that support the similarity predictions should be clearly identified.

Building transparency into the machine learning model has been a major focus of recent ML research (e.g. [2, 22]), and well-founded approaches have been proposed to attribute the model's predictions to the input features. Consider $f : \mathbb{R}^d \to \mathbb{R}$ to be some prediction model, $\boldsymbol{x} \in \mathbb{R}^d$ the point of interest, and a Taylor expansion of the prediction at some well-chosen reference point $\widetilde{\boldsymbol{x}}$. In that case, we can identify the contribution of each feature $i = 1...d$ to the prediction by the first-order terms of the expansion:

$$R_i = [\nabla f(\widetilde{\boldsymbol{x}})]_i \cdot (x_i - \widetilde{x}_i). \qquad (1)$$

While first-order terms are often suitable to explain the prediction of typical ML classifiers, similarity models are better characterized by the *interaction* between the variables of the two examples being compared. Relevant information is therefore principally contained in the second-order terms of the Taylor expansion. Denoting by $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ some similarity model and $\boldsymbol{x}$ and $\boldsymbol{x}'$ two points being fed to the similarity model, the joint contribution of features i and i' of these two points to the predicted similarity is given by:

$$R_{ii'} = [\nabla^2 s(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}')]_{ii'} \cdot (x_i - \widetilde{x}_i) \cdot (x'_{i'} - \widetilde{x}'_{i'}). \qquad (2)$$

From this mathematical starting point, we proposed a method called BiLRP [7] that more robustly extracts contributions of interacting features, and that operates by propagating the similarity score backwards, layer after layer, using purposely designed propagation rules, until the input pixels are reached. The BiLRP method is itself an extension of the LRP method [3, 20] from first-order to second-order explanations.

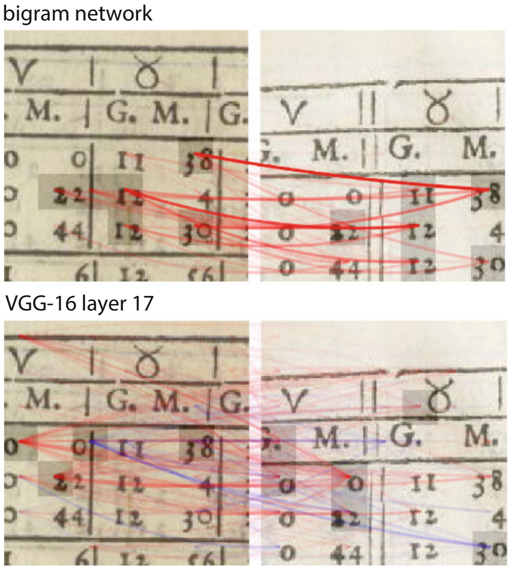

Fig. 4 Verification of the similarity models using the explanation method BiLRP. Similarity is either computed by the bigram network **(a)** or a VGG-16 model **(b)**. BiLRP explanations highlight joint feature contributions

Examples of BiLRP explanations are shown in Fig. 4 where pairs of interacting features that most strongly contribute to the similarity score are drawn with a red connection. We compare explanations of our proposed bigram network with those of a simple similarity model built on a vanilla pretrained deep CNN model for image recognition (VGG-16 [23]). For the bigram network we observe that the most relevant interacting features are indeed the shared bigrams between the pages. Since the network applies a spatial pooling over the map, we see that red connections can also go from one bigram to the same bigram at different locations, as visible for the bigram '12' in our example. In contrast, for VGG-16 we observe that the predicted similarity is grounded in task-irrelevant features such as borders or other geometric features, e.g. arcs and circles. In our case, the bigram network uses more meaningful features to arrive at its predicted similarity score, and it should therefore be preferred.

5 Connecting Illustrations Extracted from the *Sphaera* Editions

In order to investigate the evolution of the visual repertoire within the *Sphaera* corpus, and with it the evolution of knowledge in the early modern period, we also turned our attention to the more than 20,000 scientific illustrations contained within the editions of the corpus.

We start our analysis by identifying "shared illustration" groups, which we define as two or more illustrations coming from different editions that express the same semantic content iconograpically. While these groups can be primarily considered as a semantic category, the illustrations they contain usually display high visual resemblance. Shared illustrations could have been printed using the same woodblock. In other cases, printed illustrations served as blueprints for the carving of new woodblock copies. While these copies usually remained true to the original illustration design, in some instances, they introduced a degree of variation.

Computationally detectable visual resemblance is often a good proxy for the "shared illustrations," but is neither a sufficient nor a necessary criterium. Whether two illustrations are *shared* can ultimately only be decided by domain experts taking into account the context, which may necessitate reading and interpreting the accompanying text. Accordingly, we implemented the following workflow for identifying groups of shared illustrations. (1) Extraction of illustration from the corpus, (2) computational clustering illustrations according to visual similarity, (3) adjustment of calculated clusters by domain experts.

5.1 Illustration Extraction

In the first step, the illustrations were extracted from the corpus with the help of project assistants, who captured the bounding boxes of all visual content within the *Sphaera* corpus. On the one hand, this data provides the basis for the subsequent computational similarity clustering, on the other, it has provided us with the opportunity to train, in the near future, a machine learning model that facilitates the detection of illustrations in early modern period sources.

Taking advantage of the recent developments in Deep Neural Networks (DNN) for object detection, as well as their adaptability to the domain of historical documents, we trained[7] a YOLO [10] model using the manually extracted visual content bounding boxes from the *Sphaera* corpus. Our model yields an AP of 0.983 for the detection of visual elements in the corpus, and shows promising results in comparison to other models such as [19], and a generalization ability across similar datasets (e.g. Mandragore, RASM2019, IlluHisDoc).

5.2 Computational Similarity Clustering

The next step in the workflow is the noise invariant similarity illustration clustering. We initially applied an image hashing approach to abstract the illustrations [14]. While this yielded some promising results, it was not robust against some of the noise in our data, such as rotations of the illustrations originating from the original digitization of the sources. As a consequence we resorted to Convolutional Neural Networks (CNN).

In this case, we rely on the feature extracting nature CNNs, more specifically VGG16 [23], to extract representative feature maps for each of the *Sphaera* illustrations. In order to empirically test which feature maps yield the best results for our downstream task, we passed all the illustrations to different VGG16 models, cut at different layers. The resulting feature maps were then clustered using k-means with various cluster counts and compared to a limited number of hand-extracted target similarity groups. It was thus determined that the similarity clustering most suitable for the downstream task of human post-processing and cluster analysis was obtained by clustering the output of the fourth pooling layer of the VGG16 model.

While the above approach yielded acceptable results, some clusters containing a large number of heterogeneous data persisted. Thus in order to further improve our results, we resorted to image registration and pixel wise comparison. We used Oriented FAST and Rotated BRIEF (ORB) [21] to extract keypoints from binarized illustrations, which were later matched pairwise. After filtering out low quality matches, we estimated the affine transformation matrix using random sample consensus (RANSAC) to map illustration pairs onto each other. We then calculated the Structural Similarity Index Metric (SSIM) [29] between all possible pairs of registered illustrations, and detected similarity based on a predefined SSIM threshold. Finally, we created a graph with the illustrations as nodes and similarity relations as edges. This graph had a community structure, consisting of clusters densely connected internally and weakly among each other. We retrieve these community clusters, which serve as our image clusters, by means of a customized Louvain Community Detection Algorithm [5].

Direct comparison between the above proposed methods shows that the results of the latter approach are better than those achieved by clustering the CNN embeddings. Despite substantial noise in the data, this approach allows us, in the majority of cases, to distinguish illustrations made from a particular woodblock from those made by a different re-carved or copied one, as illustrated by the example in Fig. 5

5.3 Domain Experts Adjustments

The complexity of the visual elements as well as the variation in their semantic meaning necessitated the intervention of domain experts who were able to build on the results of the above computational steps. These experts cleaned

[7] Using transfer learning.

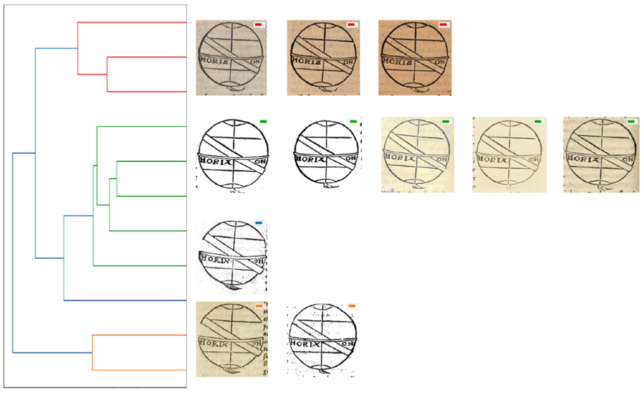

Fig. 5 For a set of similar illustrations identified by feature map clustering using VGG16, we perform a hierarchical clustering of the SSIM between the registered pairs of illustrations to retrieve clusters (color-coded in figure) containing items printed from the same woodblock. The computational result has been confirmed by a domain expert

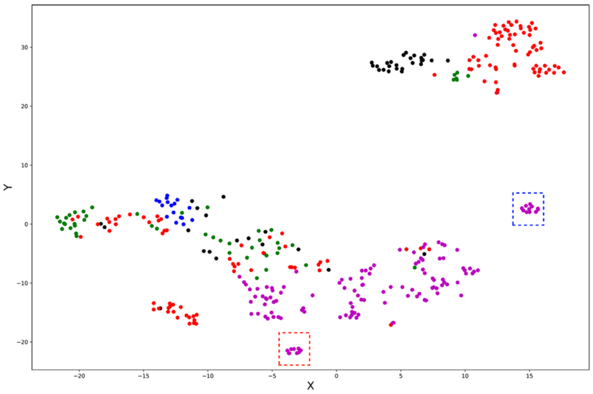

Fig. 6 t-SNE representation of the *Sphaera* edition embeddings highlighting edition types. Blue: Original texts: Red: Annotated Original Text and Compilations; Black: Compilations; Green: Annotated Original Text; Magenta: Adaptions

and grouped these clusters based on historical semantic meaning, a task that cannot be accomplished by the above computational approaches. These semantic groups are organized according to a taxonomy based on form and function. We investigate the scientific content conveyed by the images and its connections to the textual knowledge they accompany. By constructing a taxonomy and examining it in relation to the detailed information we have of each image, we aim to map how these various image characteristics changed and developed through space and time. The examination of large amount of sources reveals a different picture than the one that could be obtained through singular exemplary cases. Accordingly, we are able to achieve a new understanding of the past uses of scientific illustrations as well as of the different ways in which cosmological content was understood, applied, taught, and socially conceived in the early modern period.

6 Knowledge Graph Entity Embedding using cidoc2vec

With the increased complexity of the *Sphaera* corpus and the relatively large amount of entities used to describe each edition, it has become more challenging to describe each edition using the multitude of knowledge atoms connected to it. However, the need for such a descriptive feature vector grows increasingly as it allows us to better characterize vertices in future network analyses, where we aim to analyze attributed networks.

With this goal in mind, we developed an approach that caters to the special structure of CIDOC-CRM knowledge graphs. Each edition in the *Sphaera* knowledge graph is described by a relatively large network of entities and attributes representing different knowledge atoms. This highly atomized knowledge graph structure means that traditional knowledge graph embedding methods relying on triples facts are inefficient. This inefficiency stems from the fact that within the *Spheara* corpus knowledge graph we are often dealing with chains of relations, rather than relation triples. This is exemplified by the relation between the book edition and its contained parts using the CIDOC-CRM terminology[8]:

> sphaeara:Book – **P128:carries** – F24:Publication – **P165:incorporates** – F22:Self-Contained Expression – **P148:has component** – sphaera:Part

In order to generate meaningful entity embeddings, and keeping true to humanities research objectives which rarely require link prediction but instead focusing on data exploration and recommendation generation, we propose cidoc2vec [8], which generates entity embeddings through biased walks across a CIDOC-CRM knowledge graph. The approach is composed of two modules: a biased walk named *Relative Sentence Walk*, or RSW, and a Natural Language Processing module which relies on the well established doc2vec algorithm to generate paragraph embeddings [16].

The RSW, inspired by the syntactical similarity between knowledge graph reading and natural language, has two main objectives. The first is to collect the attributes of any main entity within the CIDOC-CRM model by 'reading' n 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. These biased walks are directed using terms calculated based on the Entity's position within the Knowledge Graph, and the amount of information it transmits based on its out-degree [8]. The n

[8] Relations are shown in bold.

walks per main entity generate a sentence document that we consider to be a representative of each entity. We then rely on doc2vec [16] to generate vector embeddings of each document.

Through the application of cidoc2vec on the *Sphaera* knowledge graph, we generated representative feature vectors whose t-SNE representation is show in Fig. 6. These feature vectors are calculated using a representative document of 500 sentences produced from each entity, and allow us to better explore our dataset to derive historical interpretations, such as the exploration of the historical reprinting phenomenon represented by very dense clusters within the blue and red box. Additionally, we can use this approach for entity alignment in CIDOC-CRM Knowledge Graphs.

7 Historical Interpretation

While the work on the knowledge atoms, such as numerical tables and scientific illustration, is still ongoing, the work concerning knowledge graph embeddings and the analysis of the behavior of the text-parts has already achieved exceptional historical results.

The first and most fundamental result of our network analysis (Sect. 3) concerns the process of homogenization of knowledge and, specifically, the mechanism that led to such an output, which we now can best describe as a mechanism of imitation. We were able to identify families of treatises characterized by their inherent text-parts similarity, while at the same time executed a strong influence – their content was imitated – on the content of other treatises produced elsewhere. By matching this analysis with the metadata, we were finally able to identify that the dominant family of treatises that gave birth to such a process was produced in the reformed Wittenberg. While the Protestant Reformation created a religious, institutional, and political division in Europe, it also created the backdrop against which scientists made their first step toward the formation of a community that, anachronistically, could be compared to a modern international scientific society. Other phenomena that could be identified in the corpus, as mentioned in Sect. 3 and that we defined as “Enduring innovations” and “Great transmitters,” show the relevance of the place “Wittenberg,” especially around the middle of the 16th century [30]. At this point, Wittenberg changed its strategy, moving from a more radically innovative position toward integrating innovations and tradition that would support the primacy of Wittenberg’s scientific literature in Europe for many decades, furnishing therefore the fuel for a long-term process of homogenization. At the end of the 16th century, based on a mix of imitation and a center-emanating output of innovations, students across Europe, from Krakow to Lisbon, were all learning the same astronomy and cosmology. These discoveries, finally, allowed us to identify the mechanism of imitation as a fundamental behavior in society to explain the process of homogenization of knowledge. We therefore investigated it in details in reference to the social networks of both authors of the commentaries and their producers, namely printers and publishers [24, 25] and found out that the main reason for such mechanism was its capacity to reduce financial risk in the framework of the then emerging and de-regulated academic book market.

8 Discussion

Specific studies, such as those dedicated to the identification of anonymous influential scholars [27], as well as studies based on distant reading approaches are still ongoing. In particular, the present goal is to develop a method for the unsupervised clustering of the numerical tables as well as for a complete evaluation of the evolution of knowledge as carried by the visual material.

Beyond implementing cross-disciplinary research that spans machine learning development, the physics of complex systems, and the history of science, we hope to establish approaches, methods, and workflows that are paradigmatic for the future development of Computational History. The project therefore also serves as an example to trace some of the main features of Computational History or, more generally, the Digital Humanities (DH).

A fundamental novelty is that digital scholarship prompts scholars to approach their research subjects differently: they have to model them. This entails, among other things, the need to make historical statements, questions, and assumptions explicit. A thorough categorization of the different text-parts or a precise idea of how images can be semantically differentiated, for instance, formed the basis of further *Sphaera* research. The decisions of *what* to make explicit (and thus to include in the research subject) and *how* to do so thus force scholars to conceptualize their subjects in very concrete terms. However, they must do so using predefined concepts, structures, and ways to relate them. In the case of the *Sphaera* project, this means thinking of the corpus as well as the related research questions in terms of triples, entities and relations, and vocabularies. By following these instructive models, historical information is expressed in ways that make it readable and operationalizable by machines. Interestingly, the act of making explicit, or what could be seen as the process of “narrowing down” meaning and dimensionality, is what enables “blowing” it up again. Only on the basis of the historical data’s initial explication according to a specific model can it then be used to form the complex networks or the “ever-expanding knowledge graph.”. Their analysis, in turn, provides research results that entirely rely on computing. In this regard, applying computing with a specific research question in mind can be seen as a product of having narrowed down the large and complex dataset, making it intelligible for humans.

The act of making historical knowledge explicit of course requires a deep understanding of the subject at hand. This is particularly evident from the fact that determining what kinds of editions exist and how exactly they differ from one another could only be achieved by manually analyzing the sources. This aspect, together with the workflow concerning the clustering of images, make another feature of digital scholarship particularly obvious: the constant interplay of human and machine, as well as the iterative workflow they engage in. “Machine” here rather generally denotes the application of physics of complex systems or the use of ML, while “human” refers to the involved scholars, or in the case of the image clustering, the “domain experts”. It is the interplay of the computing human components, the decisive factor when it comes to the success of research efforts.

Despite a dependence on financial resources, human expertise, and technological developments, the research practices that have emerged in the context of the Digital Humanities and Computational History are extending traditional approaches, developing new ones, and are at the same time increasingly engaging in productive self-reflexive debates about the approaches and methods they apply. Reflecting on research projects case-by-case, similar to what we have attempted here, may reveal common features of DH research and – in the future – make it possible to evaluate to what extent these approaches have qualitative effects on the production of our historical knowledge.

Acknowledgements This work was partly funded by the German Ministry for Education and Research (under refs 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18056A, 01IS18025A and 01IS18037A) and BBDC/BZML and BIFOLD, and by the Max Planck Institute for the History of Science. K-R.M. was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea Government (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University). We would like to thank the project assistants Beate Federau, Victoria Beyer, Salome König, Nana Citron, Gesa Funke, and Daan Lockhorst, our former colleague Olga Nicolaeva, and Lindy Divarci and Nathaniel LaCelle-Peterson for proofreading the text.

Author contribution Conceptualization H.H., M.V.; Revisions H.H., M.V.; Sect. 1 H.H., M.V.; Sect. 2 H.H., M.V.; Sect. 3 M.Z., H.K.; Sect. 4 O.E., G.M., K-R.M.; Sect. 5 J.B., J.M., N.S.; Sect. 6 H.H., M.V.; Sect. 7 M.V., H.H.; Sect. 8 A.S.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Adam K, Al-Maadeed S, Akbari Y (2022) Hierarchical fusion using subsets of multi-features for historical arabic manuscript dating. J Imaging. https://doi.org/10.3390/jimaging8030060
2. Arrieta AB, Rodríguez ND, Ser JD, Bennetot A, Tabik S, Barbado A, García S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115
3. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. Plos One 10(7):e130140
4. Bekiari C, Bruseke G, Doerr M, Ore CE, Stead S, Velios A (2021) Definition of the cidoc conceptual reference model v7.1.1. The CIDOC conceptual reference model special interest group https://doi.org/10.26225/FDZH-X261
5. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech (10):P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
6. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics, pp 4171–4186 https://doi.org/10.18653/v1/N19-1423
7. Eberle O, Büttner J, Kräutli F, Müller KR, Valleriani M, Montavon G (2020) Building and interpreting deep similarity models. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2020.3020738
8. El-Hajj H, Valleriani M (2021) Cidoc2vec: Extracting information from atomized cidoc-crm humanities knowledge graphs. Information. https://doi.org/10.3390/info12120503
9. Görz G, Seidl C, Thiering M (2021) Linked biondo: modelling geographical features in renaissance texts and maps. E Perimetron Int Web J Sci Technol Affined To Hist Cartogr Maps 16(2):78–93
10. Jocher G, Stoken A, Chaurasia A, Borovec J, NanoCode012, TaoXie, Kwon Y, Michael K, Changyu L, Fang J, V A, Laughing, tkianai, yxNONG, Skalski P, Hogan A, Nadar J, imyhxy, Mammana L, AlexWang1900, Fati C, Montes D, Hajek J, Diaconu L, Minh MT, Marc, albinxavi, fatih, oleg, wanghaoyang0106 (2021) ultralytics/yolov5: v6.0. https://doi.org/10.5281/zenodo.5563715
11. Ke Q, Ferrara E, Radicchi F, Flammini A (2015) Defining and identifying sleeping beauties in science. Proc Natl Acad Sci USA 112(24):7426–7431
12. Koho M, Ikkala E, Leskinen P, Tamper M, Tuominen J, Hyvönen E (2021) Warsampo knowledge graph: Finland in the second world war as linked open data. SW 12(2):265–278
13. Kräutli F, Valleriani M (2018) Corpus*Tracer*: a cidoc database for tracing knowledge networks. Digit Scholarsh Humanit 33(2):336–346. https://doi.org/10.1093/llc/fqx047
14. Kräutli F, Lockhorst D, Valleriani M (2020) Calculating sameness: Identifying early-modern image reuse outside the black box. Digit Scholarsh Humanit 36(2):165–174. https://doi.org/10.1093/llc/fqaa054
15. Kräutli F, Chen E, Valleriani M (2021) Information and knowledge organisation in digital humanities. In: chap Linked data strategies for conserving digital research outputs. Routledge, London, pp 206–224 https://doi.org/10.4324/9781003131816

16. Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: Xing EP, Jebara T (eds) Proceedings of the 31st international conference on machine learning, PMLR Bejing. vol 32, pp 1188–1196 (https://proceedings.mlr.press/v32/le14.html)
17. Lee BCG, Mears J, Jakeway E, Ferriter M, Adams C, Yarasavage N, Thomas D, Zwaard K, Weld DS (2020) The newspaper navigator dataset: Extracting headlines and visual content from 16 million historic newspaper pages in chronicling america. In: Proceedings of the 29th ACM international conference on information and knowledge management, association for computing machinery CIKM '20. New York, pp 3055–3062 https://doi.org/10.1145/3340531.3412767
18. van der Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(86):2579–2605 (http://jmlr.org/papers/v9/vandermaaten08a.html)
19. Monnier T, Aubry M (2020) docExtractor: an off-the-shelf historical document element extraction. In: ICFHR
20. Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR (2019) Layer-wise relevance propagation: an overview. In: Explainable AI. Lecture Notes in Computer Science, vol 11700, pp 193–209
21. Rublee E, Rabaud V, Konolige K, Bradski G (2011) Orb: an efficient alternative to sift or surf. In: 2011 International Conference on Computer Vision, pp 2564–2571 https://doi.org/10.1109/ICCV.2011.6126544
22. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR (2021) Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE 109(3):247–278
23. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations. ICLR,
24. Valleriani M (ed) (2020) De sphaera of Johannes de Sacrobosco in the Early Modern Period: The Authors of the Commentaries. Springer, Cham https://doi.org/10.1007/978-3-030-30833-9
25. Valleriani M, Ottone A (eds) (2022) Publishing Sacrobosco's "de Sphaera" in early modern Europe. Modes of material and scientific exchange. Springer International Publishing, Cham https://doi.org/10.1007/978-3-030-86600-6
26. Valleriani M, Kräutli F, Zamani M, Tejedor A, Sander C, Vogl M, Bertram S, Funke G, Kantz H (2019) The emergence of epistemic communities in the *Sphaera* corpus: Mechanisms of knowledge evolution. J Hist Netw Res 3:50–91. https://doi.org/10.25517/jhnr.v3i1.63
27. Valleriani M, Federau B, Nicolaeva O (2022) The hidden praeceptor: how Georg Rheticus taught geocentric cosmology to Europe. Perspect Sci 30(3). https://doi.org/10.1162/posc_a_00421
28. van Ingeborg V (2017) Using multi-layered networks to disclose books in the republic of letters. J Hist Netw Res 1(1):25–51. https://doi.org/10.5072/jhnr.v1i1.7
29. Wang Z, Bovik A, Sheikh H, Simoncelli E (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861
30. Zamani M, Tejedor A, Vogl M, Kräutli F, Valleriani M, Kantz H (2020) Evolution and transformation of early modern cosmological knowledge: a network study. Sci Rep. https://doi.org/10.1038/s41598-020-76916-3