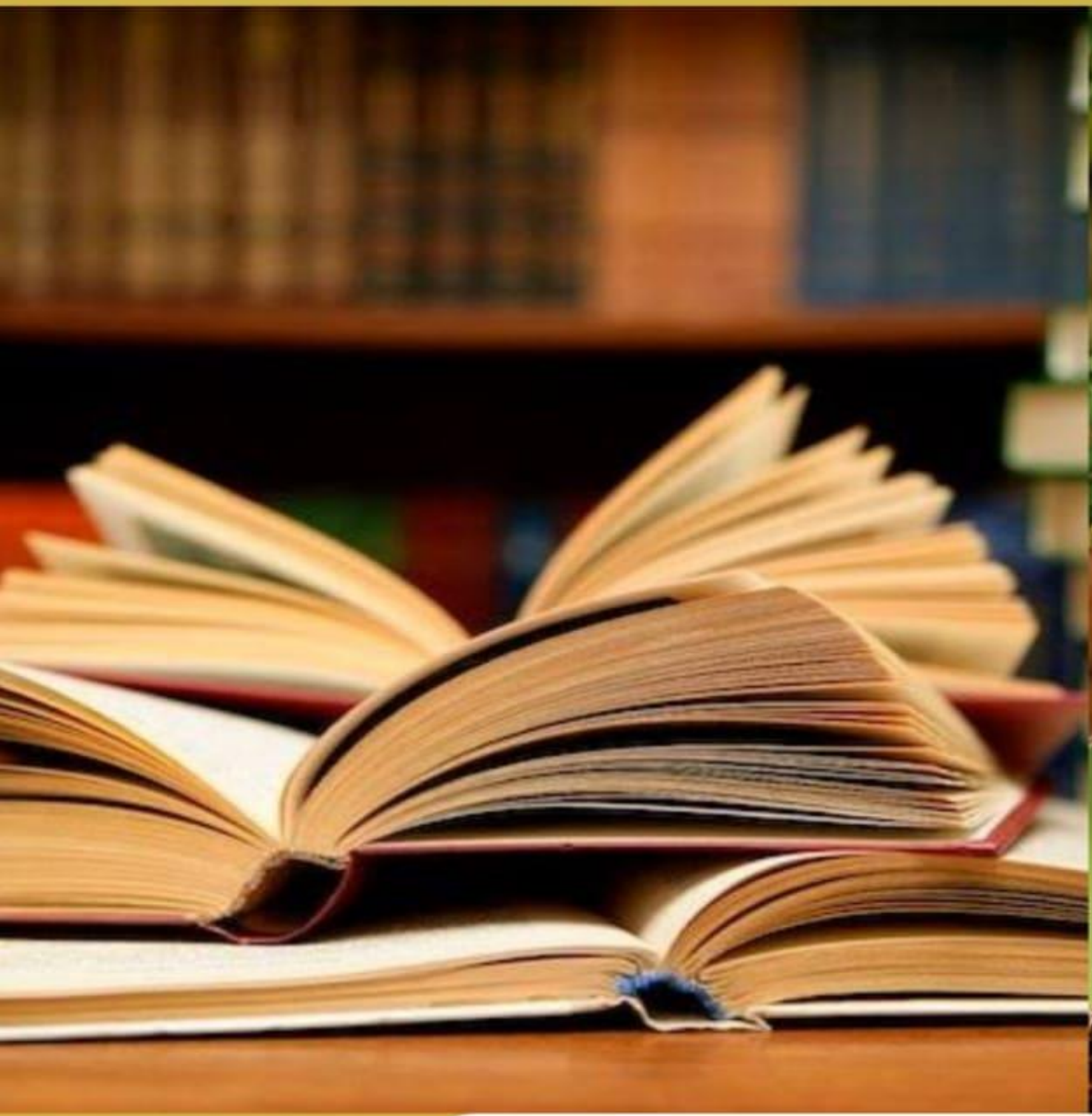
Journal of Language and Linguistics

(An open access peer reviewed Journal)



SIR PUBLISHERS

www.sirpublishers.com

ARTICLES:

INTRODUCTION TO COLLABORATIONS IN RESEARCH INFRASTRUCTURES

Author: **Jennifer Edmond**

KEYWORD(S):
Open Science; Research Infrastructures; managing and sharing research data data management planning; RDM; DMP; digital humanities; collaboration; interdisciplinary; interdisciplinarity

Text for index page:

PARTHENOS

Archived snapshot of the Introduction to Collaborations in Research Infrastructures module, which is part of the PARTHENOS Training suite [1], which was developed as part of Work Package 7 in the PARTHENOS project [2].

By the end of this module, learners should be able to:

- Understand what is meant by collaboration in humanities research
- Be aware of how this model impacts upon the development of digital humanities, and digital humanities research infrastructures

Background:

The PARTHENOS project [3] recognised that over the past ten years, researchers, institutional leaders and policymakers have begun to speak more and more about infrastructure. As more voices join the conversation, however, it can sometimes become more difficult, rather than less, to understand what exactly research infrastructure is and does. In particular in the humanities, and the digital humanities, the term is used to cover a lot of different projects, resources and approaches.

To address this gap, the PARTHENOS cluster of humanities research infrastructure projects devised a series of training modules and resources for researchers, educators, managers, and policy makers who want to learn more about research infrastructures and the issues and methods around them.

The modules, which have been released on a rolling basis from late 2016, cover a wide range of awareness levels, requirements and topic areas within the landscape of research infrastructure.

This deposit is never intended to replace the online version of the training material on the PARTHENOS website, and is intended as an archive of content.

[1] https://training.parthenos-project.eu/

[2] WP7 – Skills, Professional Development and Advancement: http://www.parthenos-project.eu/resources/projects-deliverables#1523355756261-be477222-2866

[3] http://www.parthenos-project.eu/

[This is an archived snapshot of an online course. The online course may be updated over time, and though new versions will be created to reflect major changes, the archived version may not match exactly the content of the online version]

TEXT ANALYSIS USING DEEP NEURAL NETWORKS IN DIGITAL HUMANITIES AND INFORMATION SCIENCE

Author: **Kristina Edmond**

ABSTRACT:

Combining computational tehischnologies and humanities is an ongoing effort aimed at making resources such as texts, images, audio, video, and other artifacts digitally available, searchable, and analyzable. In recent years, deep neural networks (DNN) dominate the field of automatic text analysis and natural language processing (NLP), in some cases presenting a super-human performance. DNNs are the state-of-the-art machine learning algorithms solving many NLP tasks that are relevant for Digital Humanities (DH) research, such as spell checking, language detection, entity extraction, author detection, question answering, and other tasks. These supervised algorithms learn patterns from a large number of "right" and "wrong" examples and apply them to new examples. However, using DNNs for analyzing the text resources in DH research presents two main challenges: (un)availability of training data and a need for domain adaptation. This paper explores these challenges by analyzing multiple use-cases of DH studies in recent literature and their possible solutions and lays out a practical decision model for DH experts for when and how to choose the appropriate deep learning approaches for their research. Moreover, in this paper, we aim to raise awareness of the benefits of utilizing deep learning models in the DH community.

INTRODUCTION

The research space of digital humanities (DH) applies various methods of computational data analysis to conduct multi-disciplinary research in archaeology (Eiteljorg, 2004; Forte, 2015), history (Thomas, 2004; Zaagsma, 2013), lexicography (Wooldridge, 2004), linguistics (Hajic, 2004), literary studies (Rommel, 2004), performing arts (Saltz, 2004), philosophy (Ess, 2004), music (Burgoyne, Fujinaga, & Downie 2015; Wang, Luo, Wang, & Xing, 2016), religion (Hutchings, 2015) and other fields. The scope of DH continues to expand with the development of new information technologies, and its boundaries remain amorphous (McCarty, 2013). Therefore, DH's definition is unclear and may have different interpretations (Ramsay, 2016; Poole, 2017). Library and Information Science (LIS) and DH research have a similar and overlapping scope and interfaces (Posner, 2013; Koltay 2016), to the extent that some propose to integrate and combine both research fields (Sula, 2013; Robinson, Priego, & Bawden, 2015). DH and LIS academic units are often located together (Sula, 2013), and share a significant volume of common topics, such as metadata, linked data and ontologies, information retrieval, collection classification, management, archiving and curation, bibliographic catalogue research, digitization of printed or physical artifacts, preservation of cultural heritage, data mining and visualization, and bibliometrics (Svensson, 2010; Russell 2011; Gold, 2012; Warwick 2012; Sula, 2012; Beaudoin, & Buchanan, 2012; Sula 2013; Drucker, Kim, Salehian, Bushong, 2014; Koltay, 2016; Gold, & Klein 2016). However, regardless of the definition or research scope, many (if not most) of the

research in DH/LIS focuses on textual resources, recorded information, and documents (Robinson et al., 2015; Poole, 2017). Therefore, this paper argues that a deep understanding of text analysis methods is a fundamental skill that future (and present) DH/LIS experts must acquire.

Supervised deep neural networks (deep learning) are a subset of machine learning algorithms considered to be the state-of-the-art approach for many NLP tasks, such as entity recognition (Li, Sun, Han, & Li, 2020), machine translation (Yang, Wang, & Chu, 2020), part-of-speech tagging and other tasks (Collobert & Weston, 2008) from which many DH/LIS text analysis research projects can benefit. Therefore, this paper aims to raise the awareness of DH and LIS researchers of state-of-the-art text analysis (NLP using deep neural networks) approaches and techniques. This is not the first attempt to make NLP technologies accessible or highlight the benefits of NLP to the DH/LIS research community (Biemann, Crane, Fellbaum, & Mehler, 2014; Kuhn, 2019; Hinrichs, Hinrichs, Kübler, & Trippel, 2019; McGillivray, Poibeau, & Ruiz Fabo, 2020). However, this paper argues that in addition to bridging between the NLP community and the DH/LIS research community, the DH/LIS research community should cultivate experts with a deep understanding of the technological space, experts that are capable of customizing and developing the technology themselves. Use of "off the shelf" tools and algorithms is no longer sustainable (Kuhn, 2019); the future DH expert must be comfortable using and adapting state-of-the-art NLP methodologies and technologies to the DH-specific tasks. To the best of our knowledge, this is the first attempt to highlight the challenges and analyze the potential solutions of the common usage of deep neural networks for text analysis in the DH/LIS space.

DNN models are often developed by computer scientists and trained, tested, and optimized for generic, open-domain tasks or by commercial enterprises for modern texts (Krapivin, Autaeu, & Marchese, 2009; Rajpurkar, Zhang, Lopyrev, & Liang, 2016). However, applying these DNN models for DH/LIS tasks and textual resources is not straightforward and requires further investigation. This paper presents the practical challenges that DH/LIS experts may encounter when applying DNN models in their research by examining multiple use cases presented in current literature, alongside an overview of the possible solutions, including deep learning technology. Although there might be other methodological challenges (Kuhn, 2019), this paper focuses on the two main practical challenges faced when applying deep learning for almost every DH research:

(1) Training data (un)availability - DH text resources are often domain-specific and niche, and contain a relatively small number of training examples; thus, there is not enough data for the DNN learning process to converge. Even when there is a large DH text corpus, there are no balanced ground truth labeled datasets (i.e., datasets with the distribution of "right" and "wrong" examples representative of the corpus) from which the DNN can learn (McGillivray et al., 2020), and changes or adaptations in the network architecture are required in order to achieve high accuracy for such datasets (Hellrich & Hahn, 2016).

(2) Domain adaptation - in many tasks considered "common" in NLP, the DH interpretation of the task is different from the standard interpretation. Moreover, DH text resources may need to be preprocessed before serving as input to DNNs, due to "noisy" data (biased, contains errors or missing labels or data (Hall, 2020; Prebor et al., 2018)) or non-standard data structure, such as mixed data

formats (combining unstructured text, semi-structured and structured data in the same resource). In many cases, these resources are unsuitable for serving as an input into DNN models, or if they are used as-is, the models do not achieve maximum accuracy.

These challenges have unique implications on the utilization of DNNs with DH/LIS resources and tasks and, in various cases, may require different solutions. As a result of this study, a decision model for choosing the appropriate machine-learning approach for DH/LIS research is presented as a practical guideline for experts, with topics that digital humanitists should master being outlined.

Digital Humanities and Automatic Text Analysis

Natural Language Processing (NLP) is a research area that explores how computational techniques (algorithms) can be used to understand and transform natural language text into structured data and knowledge (Young, Hazarika, Poria, & Cambria, 2018; Chowdhary, 2020). Until a few years ago, the state-of-the-art techniques that addressed supervised natural language processing challenges were based on a mix of machine learning algorithms. NLP tasks such as text classifications, entity recognition, machine translation, and part-of-speech tagging were solved using various classic supervised machine learning algorithms, such as Support Vector Machine (SVM), Hidden Markov Model (HMM), decision trees, k-nearest neighbors (KNN), and Naive Bayes (Zhou & Su, 2002; Liu, Lv, Liu, & Shi, 2010; Vijayan, Bindu, & Parameswaran, 2017). Basically, these algorithms apply a manually selected set of characteristic features to a given task and corpus, and a labeled dataset with "right" and "wrong" examples for training the optimal classifier. Given a new example of the same type, this classifier will be able to automatically predict whether or not this example belongs to the predefined category (e.g., whether a given sentence has a positive sentiment or not).

However, in many cases, it is not easy to decide what features should be used. For example, if a researcher wishes to learn to classify a text's author from the Middle Ages, she will need to use the features that represent the unique writing styles that distinguish the authors. Unfortunately, it is not easy to describe these features in terms of textual elements. Deep learning solves this central problem by automatically learning representations of features based on examples instead of using explicit predefined features (Deng & Liu, 2018). Deep learning (DL) is a sub-field of machine learning that draws its roots from the Neurocognition field (Bengio, Goodfellow, & Courville, 2017). The DL approach uses deep neural networks (DNN) models for solving a variety of Artificial Intelligence tasks. The technical details of various DNN models and techniques appear in Appendix I.

DH researchers use NLP algorithms for DH-specific tasks in various domains. For example, Niculae, Zampieri, Dinu, and Ciobanu (2014) used NLP techniques to automatically date a text corpus. They developed a classifier for ranking temporal texts and dating of texts using a machine learning approach based on logistic regression on three historical corpora: the corpus of Late Modern English texts (de Smet, 2005), a Portuguese historical corpus (Zampieri & Becker, 2013) and a Romanian historical corpus (Ciobanu, Dinu, Dinu, Niculae, & Sulea, 2013). To construct social networks among literary characters and historical figures, Elson, Dames, and McKeown (2010) applied "off-the-shelf" machine learning tools for natural language processing and text-based rules on 60 nineteenth-century British novels. Zhitomirsky-Geffet and Prebor (2019) used lexical patterns for Jewish sages

disambiguation in the Mishna, and then applied several machine learning methods based on Habernal and Gurevych's (2017) approach for the co-occurrence of sages and pattern-based rules for specific inter-relationship identification in order to formulate a Jewish sages social interactions network. In paleography, the study of historical writing systems and the deciphering and dating of historical manuscripts, Cilia, De Stefano, Fontanella, Marrocco, Molinara, and Freca (2020) utilized MS-COCO (Lin, Maire, Belongie, Hays, Perona, Ramanan, & Zitnick, 2014), a generic corpus of images, and a domain-specific corpus to train DNN models and design a pipeline for medieval writer identification. To predict migration and location of manuscripts, Prebor, Zhitomirsky-Geffet and Miller (2020a, 2020b) devised lexical patterns for disambiguation of named entities (dates and places) in the corpus of the Department of Manuscripts and the Institute of Microfilmed Hebrew Manuscripts in the National Library of Israel. Next, the authors trained a CART machine learning classifier (Classification and regression tree based on Decision Tree learning) (Rokach and Maimon, 2015) to predict the places of manuscripts that were often absent in the corpus. For ancient languages analysis, a study (Dereza, 2018) compared accuracy for lemmatization for early Irish data using a rule-based approach and DNN models, and proved the advantages of using DNN on such a historical language - even with limited data. For historical network analysis, Finegold, Otis, Shalizi, Shore, Wang, and Warren (2016) used named entity recognition tools (Finkel, Grenager, & Manning, 2005; Alias-i, 2008) with manual rules on the Oxford Dictionary of National Biography and then applied a regression method, namely Poisson Graphical Lasso (Yang, Ravikumar, Allen & Liu, 2013) to find correlations between entities (nodes). Nevertheless, as demonstrated by the examples above, although there is a "computational turn" (Berry, 2011) in the DH research and methodologies, state-of-the-art computational NLP algorithms, like deep neural networks, are still rarely used within the core research area of DH (Kuhn, 2019).

To estimate the potential of deep learning use in DH, a comparison has been performed to one of the fields that is similar to DH - Bioinformatics. These fields are comparable since both are characterized by their inter-disciplinarity and because Bioinformatics thrives on application of computational analysis for exploring and investigating information repositories in a chosen knowledge domain (Ewens & Grant, 2006). A list of leading journals was compiled in each field and searched for articles with "deep neural network" and "machine learning" keywords. For DH, twelve journals were selected, based on Spinaci, Gianmarco, Colavizza, Giovanni, & Peroni (2019), all in English and ranked as 1 (exclusively DH). For Bioinformatics, twelve journals were selected based on Google Scholar's top publication list[1]. The two lists of the journals appear in Appendix III.

The comparison was conducted on the articles published in the above journals over the past three years and measured the following: 1) the percentage of articles with each of the two keywords in the selected journals in each field, to ascertain the usage of machine learning (ML) in general vs. deep learning (DL) in particular, in each field; and 2) the percentage of articles mentioning deep learning out of the machine learning articles in each field. As can be observed from Figure 1, in the DH field,

[1] https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=bio_bioinformatics

only 21% of the articles discussing "machine learning" also discussed "deep learning"; while in Bioinformatics, 52% of the articles discussing "machine learning" also discussed "deep learning". Moreover, in the DH field, only 3.8% of the articles mentioned "deep learning", while in Bioinformatics, 19.5% of the articles mentioned "deep learning" – five times higher. In addition, in the DH field, 18% of the articles discussed "machine learning", while in Bioinformatics, 37% of the articles discussed "machine learning" – only two times higher. These results indicate that the DH field "lags behind" when it comes to using machine learning and especially deep learning state-of-the-art models.

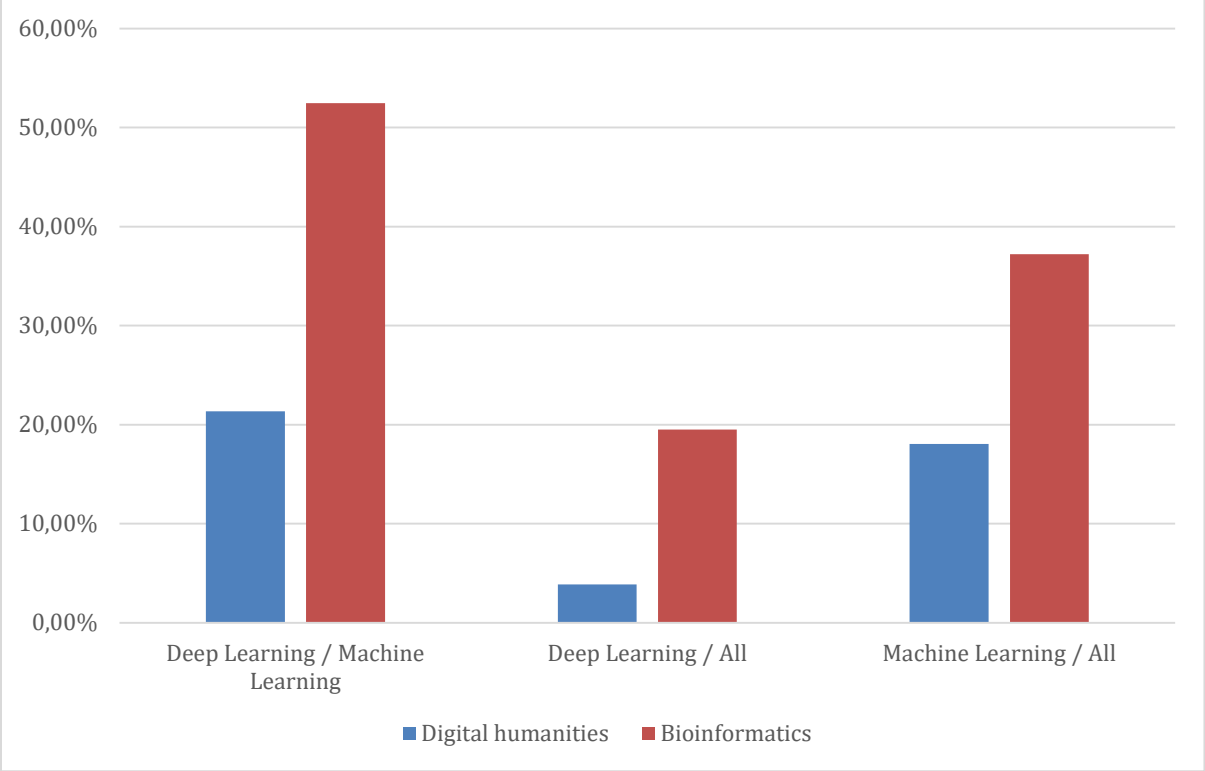


Figure 1: Deep neural networks and machine learning articles in DH/LIS vs. Bioinformatics.

The next section provides an in-depth analysis of challenges and potential solutions for using DNN in DH/LIS, supported by multiple use-case studies from the recent DH literature. The analysis is divided into two main sections dealing with two primary challenges in applying deep learning to DH research: training data (un)availability and domain adaptation.

Challenges when Using Deep Learning for Digital Humanities Research

Training Data (Un)availability

Computer scientists often work on generic supervised text analysis tasks with open-domain or modern datasets. Kaggle[2], the machine learning community, hosts many of these datasets. For example, the IMDb dataset contains a short description of a movie, and its review score allows to research sentiment analysis (Maas, Daly, Pham, Huang, Ng, & Potts, 2011); question answering system can be developed using Stanford Question Answering Dataset (Rajpurkar, Jia, & Liang, 2018);

[2] https://www.kaggle.com/datasets

and SPAM filtering can be developed using a dedicated dataset (Almeida, Hidalgo, & Silva, 2013). Unfortunately, the DH community has not (as yet) produced large annotated open datasets for researches (although there are few in niche areas like (Rubinstein, 2019; Chen & Chang, 2019)). The lack of annotated data is a challenge for both classical machine learning and deep learning supervised algorithms (Elmalech & Dishi 2021). However, supervised deep learning algorithms require significantly more data than machine learning algorithms, making this challenge a critical practical challenge for DH researchers. This is one reason that even when DH/LIS researchers use deep learning, they often use unsupervised algorithms that do not require training data and are limited to specific tasks (Moreno-Ortiz, 2017). This section investigates some of the methods that DH researchers can apply to overcome this challenge.

Training Dataset Generation by Humans

Humans are the best alternative for dataset generations due to their domain knowledge and high accuracy. Therefore, the first consideration when generating a dataset is to consider if humans can be used for the job. However, humans are not as scaleable as computer software. It is possible to manually generate a dataset by humans when the needed labeling is relatively small or as a baseline for synthetic dataset generation. There are two types of manual dataset generation: crowd-based dataset generation and domain expert-based dataset generation. Crowdsourcing dataset generation is a relatively cheaper and effective method, but it can only be used when the labeling is "common knowledge". In some cases, for example, in the study aiming to generate a dataset of relationships extraction between characters in literary novels (Chaturvedi et al., 2016), the researchers must use expert annotators that can read and understand a novel; or even annotate themselves when working with historical languages known only to a few, as in Schulz & Ketschik (2019).

Crowdsourcing is based on large groups of non-expert, low-paid workers or volunteers performing various well-defined tasks. Existing studies tested optimization strategies for different tasks, such as extracting keyphrases (Yang, Bansal, Dakka, Ipeirotis, Koudas, & Papadias, 2009), natural language and image annotation (Snow, O'Connor, Jurafsky, & Ng, 2008; Sorokin & Forsyth, 2008), and document summarization (Aker, El-Haj, Albakour, & Kruschwitz, 2012). Crowdsourcing requires quality control to ensure that crowd workers are performing their tasks at a satisfactory level (Elmalech Grosz 2017). One of the effective generic (task-agnostic) quality control techniques is filtering out tasks with a low inter-worker agreement (Bernstein, Little, Miller, Hartmann, Ackerman, Karger, Crowell, & Panovich, 2010; Downs, Holbrook, Sheng, & Cranor, 2010; Kittur, Smus, Khamkar, & Kraut, 2011). Another popular approach is breaking tasks into sub-tasks (Bernstein et al., 2010; Kittur et al., 2011).

Employing crowd workers for dataset generation has been carried out in various domains, including DH projects (e.g., Elson, Dames, & McKeown, 2010). Thus, in this use-case study, Elson et al. (2010) utilized crowdsourcing to build a dataset of quoted speech attributions in historical books in order to generate a social network among literary characters. Elson et al. (2010) did not use DNN, but rather classic machine learning methods (Davis, Elson, & Klavans, 2003), but the dataset generating process is the same for classic ML and DL.

Another example of such a use-case is fixing Optical Character Recognition (OCR) errors in historical texts. In the DH/LIS space, there is great interest in investigating historical archives. Therefore, over the past few decades, archives of paper-based historical documents have undergone digitization using OCR technology. OCR algorithms convert scanned images of printed textual content into machine-readable text. The quality of the OCRed text is a critical component for the preservation of historical and cultural heritage. Unsatisfactory OCR quality means that the text will not be searchable, analyzable, or analysis may result in wrong conclusions. Unfortunately, while generic OCR techniques and tools achieve good results on modern texts, they are not accurate enough when applied to historical texts. Post-correction of digitized small scale or niche language historical archive is a challenge that can be solved using DNNs with high accuracy (Chiron, Doucet, Coustaty, & Moreux, 2017; Rigaud, Doucet, Coustaty, & Moreux, 2019) if an appropriate dataset is attainable. Therefore, the first thing that should be researched is an effective methodology for crowdsourcing this specific task (Suissa, Elmalech, & Zhitomirsky-Geffet, 2019). The details of the crowdsourcing research are outside the scope of this paper. What is essential from the DH/LIS research point of view is that the findings of Suissa et al. (2019) proved to be an effective dataset generation approach. Using the developed strategies, DH researchers can optimize the process to achieve better results matching their objectives and priorities. The corrected corpus of OCRed texts created by the optimized crowdsourcing procedure can serve as a training dataset for DNN algorithms.

However, although the crowdsourcing method yields satisfactory results, it is suitable mainly for widely spread languages like English or Spanish. Other national languages do not have enough crowd workers-speakers to utilize such an approach effectively. Moreover, manually generating a dataset for training a DNN model in order to post-correct OCR errors is expensive and inefficient, even when the task is crowdsourced. Therefore, in practice, this human-only dataset generation should be shifted to a human-in-the-loop solution.

Training Dataset Generation using Algorithms

The next range of solutions takes a two-phase approach. In the first phase, humans are used to create a small set of examples; this set of examples is used in the second phase by a different set of algorithms to generate a synthetic dataset with numerous training examples (Pantel, & Pennacchiotti, 2006; Bunescu, & Mooney, 2007). One way is to find recurring patterns in a small number of manually corrected examples, and use them to generate more correct examples. Thus, the use-case study that adopted this approach for automatic training dataset generation in the OCR post-correction domain, Suissa, Elmalech, & Zhitomirsky-Geffet (2020) used crowd workers to fix a relatively small set of OCRed documents. Then, the Needleman–Wunsch alignment algorithm (Needleman, & Wunsch, 1970) was used to find common confusions between characters committed by the crowd workers. Using this confusion list, a large dataset of "wrong" and "right" sentences was generated and used by a DNN to correct historical OCRed text.

Another way to generate a dataset from a small set of manual examples is called "distant supervision" (Mintz, Bills, Snow, & Jurafsky, 2009). In this approach, a classifier is trained on a small set of examples and is applied to a large corpus. The classifier will classify the data with a relatively low

accuracy but sufficiently high accuracy for the DNN to learn other features from this weak classification. Blanke, Bryant, & Hedges (2020) used this method to perform sentiment analysis on Holocaust testimonials data (Thompson, 2017). In the first phase, they did not use crowd workers for the initial dataset generation but rather applied a dictionary-based approach to find negative and positive sentiment sentences based on the TF-IDF measure (Singhal, 2001). Using these sentences, they trained a classifier to distinguish between positive and negative examples. In the second phase, they used the classifier to produce a large training corpus of positive and negative memories of Holocaust survivors for DNN text analysis. Using this method eliminates the need for humans; however, it is suitable only for specific tasks.

A different approach to solving the training dataset's unavailability is the transfer learning (Torrey, & Shavlik, 2010) method. In transfer learning, a generic dataset is used; the dataset should be suitable for the task needed to be solved, but with open-domain / other domain data. The model is then trained again using a small set of domain-specific examples (generated by humans or artificially). This approach is based on the intuition that humans transfer their knowledge between tasks based on previous experiences. Cilia et al. (2020) utilized transfer learning to identify medieval writers from scanned images. Instead of generating a large dataset, they used a model that was already trained on an open generic dataset MS-COCO (Lin et al., 2014) and trained it again using a small set of domain-specific examples from the Avila Bible (images of a giant Latin copy of the Bible). Banar, Lasaracina, Daelemans, & Kestemont (2020) applied transfer learning to train neural machine translation between French and Dutch on digital heritage collections. They trained several DNNs on Eubookshop (Skadiņš, Tiedemann, Rozis, & Deksne, 2014), a French-Dutch aligned corpus. Then, instead of training the DNN models directly on the target domain data, they first trained the models on "intermediate" data from Wikipedia (articles close to the target domain). Only then did they train the models for the third time on the target domain data - the Royal Museums of Fine Arts of Belgium dataset. Using this "intermediate fine-tuning" approach, Banar et al. (2020) achieved high accuracy for French-Dutch translation in the domain of Fine Arts. This method can also solve another challenge for the DH/LIS researcher when using DNN models – the domain adaptation challenge.

Recent studies (Radford, Wu, Child, Luan, Amodei, & Sutskever, 2019; Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, & Amodei, 2020) show that in some cases, instead of fine-tuning a pre-trained model, a large-scale pre-trained model, such as GPT3 (Radford et al., 2019), trained on ~500 billion (modern) words, can achieve good results with a limited (or without) domain-specific dataset. Although these methods (named Few-shot and Zero-shot learning) do not reach the same performance as the fine-tuning method, they are preferable for low resource domains when dataset generation is impossible. However, most of the models that are pre-trained on a large-scale modern English dataset and suitable for Few-shot and Zero-shot learning may not reach the same accuracy for DH historical corpora, especially in (other than English) national languages, due to a bias towards modern language.

Domain Adaptation

Even with a large dataset ready for DNN training, there are other challenges a DH/LIS expert may encounter when attempting to solve a text analysis task on DH/LIS data with DNNs. As mentioned in the previous section, data is a critical part of DNN's high accuracy. However, specific task/domain adaptation is just as vital, and without adapting the model or the architecture to the specific task and domain, the DNN may perform poorly.

A DNN model is a set of chained mathematical formulas with weights assigned to each node (neuron) expressing a solution to a specific task. Although there are regulation techniques to generalize the DNN model, in many cases training the model with different data will significantly impact the weights. In other words, using the same mathematical formulas, the learning process interprets the same task differently. In this context, transfer learning described in the previous section can also serve as a domain adaptation method, since the DNN model's interpretation of the task is adjusted to the domain-specific data. Moreover, DH/LIS text analysis tasks are not just different in terms of interpretation but also often require domain-specific preprocessing and analysis pipeline. Therefore, in order to improve the accuracy of DNN models for text analysis tasks, DH/LIS experts should be familiar with methods and techniques for customizing DNN models, preprocessing DH/LIS data, and adapting the analysis pipeline.

DNN Optimization for DH-specific Tasks

A DNN model has a high number of architecture components and hyper-parameters that influence the model training efficacy and accuracy. Selecting the domain-specific suitable components and hyper-parameter values may considerably improve the performance of the DNN (Bengio, 2012). Here are a few of the most common architectures and hyper-parameters that an expert should consider (see Appendix I for technical details):

- Architecture components:
  - Type of the model – for instance, RNN-based, SAN-based (Vaswani et al., 2017), feed-forward-based, Transformers-based (Devlin et al., 2018).
  - Type and size of the layers – including individual layers, such as CNN (Albawi, Mohammed, & Al-Zawi, 2017), LSTM (Hochreiter et al., 1997), GRU (Cho et al., 2014), ResNet (He, Zhang, Ren, & Sun, 2016), AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), and multi-layer architectures, such as BERT (Devlin et al., 2018). These can be applied with or without bidirectionality (Schuster et al., 1997), attention (Bahdanau, Cho, & Bengio, 2015), skip-connections (Chang, Zhang, Han, Yu, Guo, Tan, & Huang, 2017), and other architectural components.
  - Type of input - DNN input is a vector (a series of numbers). Each number can represent a word using word-embedding methods, such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), a single character using one-hot encoding or character-embedding (Char2Vec), encoded features, or contextual embeddings (e.g., BERT (Devlin et al., 2018), RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, & Stoyanov, 2019), XLNet (Yang, Dai, Yang, Carbonell, Salakhutdinov, & Le, 2019)) based on the surrounding words.

- Number of layers and the DNN information flow – for instance, encoder-decoder architecture (Cho et al., 2014).
- Activation functions (the neuron's function) – including Sigmoid, Tan-h, ReLU, and Softmax.
- Loss functions (the "size" of the training error) – regression tasks can be: i) mean squared error (MSE), ii) mean squared logarithmic error, iii) mean absolute error; for binary classification tasks: i) binary cross-entropy, ii) hinge, iii) squared hinge; for multi-class classification: i) multi-class cross-entropy, ii) sparse multi-class cross-entropy, iii) Kullback- Leibler divergence.

- Hyper-parameters:
  - Type and size of the regulation layers – regulation layers reduce overfitting by adding constraints to the DNN. These constraints, such as dropout (Srivastava et al., 2014), L1, and L2, prevent the model from learning the training data and force it to learn the patterns in the data.
  - Batch size - the number of examples to use in a single training pass.
  - Number of epochs and the epochs' size - the number of iterations on the training data and the number of examples to use during the entire training process.
  - Learning rate, method, and configuration - such as stochastic gradient descent (SGD), adaptive moment estimation (Adam) (Kingma & Ba, 2014), and Adagrad (Duchi, Hazan & Singer, 2011).

Theoretically, architecture components are also hyper-parameters. However, from a practical perspective, once architecture components are chosen, they are usually fixed. There are techniques that can be applied to find and set these architecture components and hyper-parameters automatically. These techniques are called AutoML and are suitable for many different DNN models (and classical ML models). However, AutoML has its limitations: it is often costly (training the model repeatedly), does not fit large-scale problems, and may lead to overfitting (Feurer & Hutter, 2019). It is advisable to check AutoML optimization methods such as submodular optimization (Jin, Yan, Fu, Jiang, & Zhang, 2016), grid search (Montgomery, 2017), Bayesian optimization (Melis, Dyer, & Blunsom, 2017), neural architecture search (So, Liang, & Le, 2019), and others (Feurer et al., 2019) or, if the researcher has a hypothesis or intuition about the problem, it is also possible to test multiple architecture components and hyper-parameters combinations manually. Moreover, training a large DNN language model such as a BERT-based model with standard pre-defined hyper-parameters on public cloud servers costs $2,074-$12,571, depending on the hyper-parameters and the corpus size (Devlin et al., 2018), while using neural architecture search (So et al., 2019) to train a DNN language model with hyper-parameters optimized for the specified task costs $44,055–$3,201,722 (Strubell, Ganesh, & McCallum, 2019). Therefore, the budget is another consideration for using some AutoML methods.

Numerous DH studies have demonstrated the importance and the impact of hyper-parameters optimization on the DNN accuracy. Tanasescu, Kesarwani, & Inkpen (2018) optimized hyper-parameters for poetic metaphor classification. They experimented with different activation functions (ReLU, Tan-h for the inner layers and Softmax and Sigmoid for the output layer), number of layers (1-4), number of neurons in each layer (6-306), dropout rate(0-0.9), number of epochs (20-1000), and batch size (20-200). The optimization increased the metaphor classification F-score by 2.9 (from 80.4

to 83.3) and precision by 5.6 (from 69.8 to 75.4). Wang et al. (2016), used a DNN model for Chinese song iambics generation and tested several architecture components. In their research, Wang et al. (2016) added an attention layer (Bahdanau et al., 2015) on top of bidirectional LSTM layers and tested several domain-specific training methods. This DNN domain optimization made it possible to achieve near-human performance. These use-cases emphasize how important it is for DH/LIS experts to understand architecture components and hyper-parameters and their usage.

Domain-specific Dataset Adaptation for DNN

Using DNN models in some domains can also require adaptation of the data (preprocessing) prior to inputting it into the DNN model. A use-case study of Won, Murrieta-Flores, & Martins (2018) aimed to perform Named Entity Recognition (NER) on two historical corpora, Mary Hamilton Papers (modern English from 1750 to 1820) and the Samuel Hartlib collection (early modern English from 1600 to 1660). NER is an NLP task which outputs identification of entity types in text. Entity types can be places, people, or organization names and other "known names". The historical corpus selected in Won et al. (2018) was OCRed and preserved in hierarchical XML files with texts and metadata. DNN models (and the tools used in the study) for NER are not designed to work directly on XML since XML is a graph-based format, and NER is a sequence-based task. It should be noted that there are graph-based DNN models (e.g., Scarselli, Gori, Tsoi, Hagenbuchner, & Monfardini, 2008), but they are not suitable for the NER task. Therefore, Won et al. (2018) needed to adapt their domain data by "translating" the XML markup into text sequences that a DNN model can receive as input. In this preprocessing phase, the researchers took into account the metadata that exists in the domain that was embedded in the XML file, such as authorship, dates, information about the transliteration project, corrections and suggestions made by the transliterators, and particular words and phrases annotated within the body text. Moreover, the square brackets (and their content) added by the transcribers were semi-automatically removed from the text. The metadata was added to the text sequence as labels for the training data to improve the accuracy of the results. Won et al. (2018) did not use DNN models directly but rather used “off the shelf” software to conduct their research. However, they concluded the research with the recognition that using pre-made tools is not sufficient - “*Finally, it must be noted that although this research accomplished the evaluation of the performance of these NER tools, further research is needed to deeply understand how the underlying models work with historical corpora and how they differ.*”

DNN Pipeline Adaptation

DNN models are designed to work in a certain pipeline of components to solve a specific task. For example, a "naïve" DNN based pipeline for the OCR of a book collection will be: 1) scan a book page, 2) use the image as an input to an image-to-text DNN model, 3) use the obtained text or post-process it to correct errors. However, in some cases, it is advisable to design a new domain-specific pipeline to solve the task or increase the model's accuracy. A use-case of such a domain-specific OCR pipeline is presented by Cilia et al. (2020). The goal of the study was identification of the page’s writer for each page of the given medieval manuscript. Medieval handwritten manuscripts present two unique challenges for OCR: 1) first section letters or titles may be drawn as a picture over several lines, and 2)

handwritten lines are not always aligned and may reduce accuracy when performing a full-page OCR. Cilia et al. (2020) designed a pipeline for processing handwritten medieval texts with three main steps, using: 1) an object detector to detect lines in the page's scanned image and separate a picture at the top from the text lines, 2) a separate DNN classifier to classify each line, and 3) a majority vote among multiple DNN classifiers obtained for each line and picture object at the line-level, in order to make a decision for the classification (writer identification) of the entire page. This pipeline, tailored to the medieval paleography domain, solved the domain's unique challenges by separating between picture objects and text lines and classifying each line with a different classifier instead of classifying an entire page with a single DNN model (the naïve pipeline). This pipeline's domain adaptation approach combined with the transfer learning approach, described in the previous section, produced an impressive 96% accuracy in identifying writers that would not have been achieved without this adaptation.

Pipeline adaptation is not just pipelining different models or combining ML and DL; it is also re-training and adapting an existing model, i.e., fine-tuning a model. Fine-tuning a model is a subset of transfer learning, in which a model is trained on a different dataset and also changed by setting different hyper-parameters or adding new last layers on top of the model to fit a specific task. In their research, Todorov and Colavizza (2020), fine-tuned a BERT-based model (Devlin et al., 2018) for increasing the annotation accuracy of NER in French and German historical corpora. In particular, the Groningen Meaning Bank's Corpus Annotated for NER was applied (Bos, Basile, Evang, Venhuizen, & Bjerva, 2017). To embed words (including sub-words) and characters, four models were applied: (1) newly trained word-embeddings on their historical corpus, (2) in-domain pre-trained embeddings that were trained on another corpus in the same domain, (3) BERT-based embedding that was trained on French and German Wikipedia, and (4) character level embeddings learned from the historical corpus training data. As can be observed from Figure 2, Todorov et al. (2020) combined the embedding (by concatenation) and transferred the unified embeddings to a new layer based on a Bi-LSTM-CRF layer. A Bi-LSTM-CRF layer is a Bidirectional (Schuster et al., 1997) Long Short-Term Memory (Hochreiter et al., 1997) layer that merges the sub-word embedding input into a word-level output and transfers its output to fully connected layers (one layer per each entity type) which then outputs tag (entity type) probabilities for each token using Conditional Random Fields (Lafferty, McCallum, & Pereira, 2001). The Bi-LSTM-CRF method has been shown as useful and accurate by Lample, Ballesteros, Subramanian, Kawakami, & Dyer (2016). They also changed the LSTM activation function (remove the tan-h function) and tried three different hyper-parameters configurations. Using the domain-specific pipeline, model, and hyper-parameters, the researchers dramatically increase the accuracy (in some entity types by over 20%) of NER task on French and German historical corpora compared to a state-of-the-art baseline model. Moreover, they tested the impact of the pre-trained generic embedding. They found that (1) without using the open-domain embedding (BERT), their model did not attain high accuracy, and (2) on the other hand, "freezing" the open-domain embedding layers (i.e., using them but re-training only the top layers on the domain-specific historical data) did not affect the accuracy. These findings demonstrate the importance of adapting DNN models to a specific domain and task,

while reducing the training time and costs by freezing the large open-domain layers. It is essential to note that besides inputting the historical corpora documents into the DNN model, Todorov et al. (2020) also tested the addition of manually-created features to the documents such as title, numeric and other markups; these features did not have any effect on the accuracy, proving that the DNN model "learned" (or at least did not need) these features.

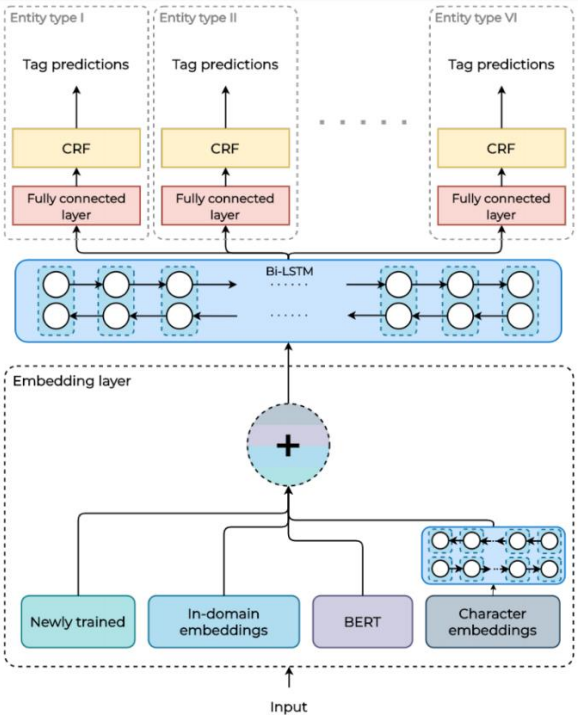


Figure 2: Historical corpora NER fine-tuning pipeline (Todorov & Colavizza, 2020).

A Decision Model for Using Deep Learning for Digital Humanities Research

Based on the above analysis of challenges and possible solutions illustrated by multiple use-case studies described in the recent literature, it is clear that the DH/LIS experts must know just enough math, understand the inner-working of ML and DL algorithms, Python programming, and use these frameworks and other popular modules (Géron, 2019).

Therefore, this paper argues that DH/LIS researchers can no longer see NLP and ML researchers as their "tool makers", and must learn to apply and adapt deep learning models (DNNs) to their specific research domain. However, since working with DNN models requires significant effort, computational resources, budget, and time, a decision model was formulated for assisting DH experts in determining when it is "worthwhile" to invest in training DNN models. The decision model is based on two strategies: 1) the data availability strategy – how to assess the types of methods and models suitable for the available dataset, and 2) the domain adaptation strategy – how to determine whether and when it is “worthwhile” to invest in domain adaptation.

Figure 3 presents the data availability strategy and leads to three possible recommendations: (1) with no data, either zero-shot DL models, or hard-coded rules/assumptions regarding domain data should be implemented, based on prior knowledge and experience; (2) with limited data, either classical

machine learning algorithms, such as SVM or HMM, or few-shot DL models can be used; otherwise (3) it is advisable to use supervised deep learning models for the task. It should be noted that if the DNN model is overfitting (high accuracy on the training dataset and low accuracy on the validation dataset), it is advisable to increase the dataset size by employing expert workers, crowdsourcing, or synthetic data generation. Figure 4 presents the domain adaptation strategy and also leads to three possible recommendations: (1) if strict rules can be defined, there is no need for ML or DL; (2) with limited resources or for low accuracy tasks, ML is the preferable option, and (3) with the appropriate resources and a need for high accuracy, DL with domain adaptation should be utilized. A researcher can use both strategies of the proposed decision model to choose the recommended approach for the given task. Since there are many different text analysis tasks, some aspects of the strategies depend on the expert's assessment; for example, "what is considered a small or a large dataset?" and "what is low or high accuracy?". These assessments should be performed by the researcher based on the concrete task, domain, and needs. Notice that the advice to use DNN models does not mean that it is not recommended to combine them with ML algorithms when suitable.

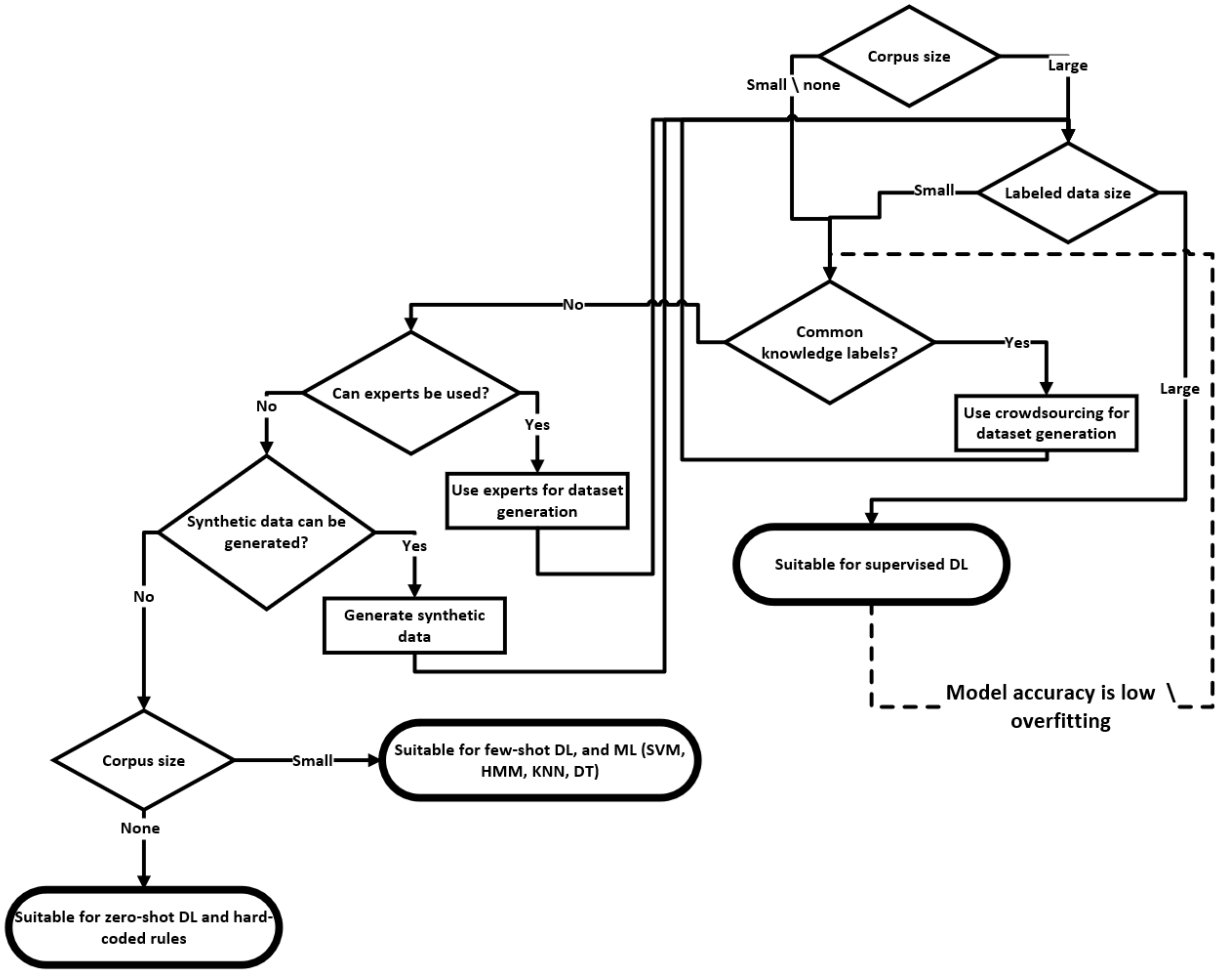


Figure 3: Data availability strategy for DH researchers

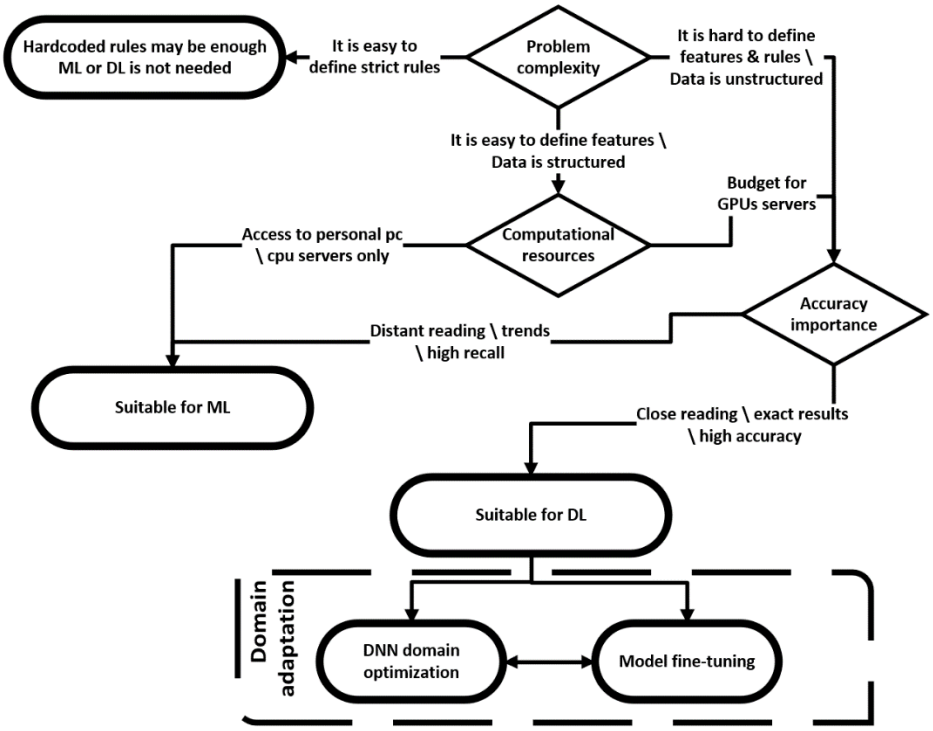


Figure 4: Domain adaptation strategy for DH researchers

As can be observed from the proposed decision model, supervised DL should be used when there is a large corpus (or a large corpus can be generated), for complex problems such as unstructured texts, when the researcher has a budget for computational resources (GPUs servers), and accuracy is essential (domain adaption is always assumed). Since most of the DH corpora are not labeled, dataset generation will most probably be required. When the labeling requires only "common knowledge", it is advisable to use crowdsourcing (if possible); otherwise, the researcher should consider using domain experts or automatic generating of synthetic data as explained above in this paper. A step-by-step example for decision model usage for a specific DH task can be found in Appendix II.

It should be noted that the extensive computational resources needed to train DNN models have an impact on the environment. DL may become a major contributor to climate change if the exponential growth of training more and more DNN models continues (Anthony, Kanding, & Selvan, 2020; Hsueh, 2020). It has been estimated that training one transformer model such as BERT-based (Devlin et al., 2018) will produce similar amounts of CO_2 to those of air travel of one person from NY to SF; using a neural architecture search (So et al., 2019), an AutoML method, will produce almost five times more CO_2 than an average car produces throughout its lifetime including the fuel (Strubell et al., 2019). We note that the proposed decision model does not consider environmental impact, yet researchers should be aware of this and take it into consideration.

By using this decision model as a guideline and applying the suggested solutions for the two fundamental challenges faced by many DH projects – DH-specific training dataset generation and model adaptation, DH/LIS experts can solve a variety of important tasks in the field for diverse national languages, such as 1) improving OCR post-correction (including restoring damaged text); 2)

automated ontology and knowledge graph construction for various DH domains (based on entity/category and relation extraction and NER); and 3) corpus-based stylometric analysis and profiling of DH resources (e.g., identification of an author, date, location, and sentiment of the given text or image).

Conclusion and Discussion

This paper presents the main two challenges almost every DH/LIS research can expect to encounter using DNN models in her research. Although classic learning techniques based on rules, patterns, or predefined features are no longer considered state-of-the-art in many text processing tasks (e.g., Thyaharajan, Sampath, Durairaj, & Krishnamoorthy, 2020; Glazkova, 2020), DH/LIS researchers are still using them often, even when there is a better alternative such as deep neural networks. The reasons for avoiding using deep learning in DH may be the lack of "off-the-shelf" tools tailored for the specified task, lack of training data, as well as time, computational resources, and budget limitations. Based on the presented investigation of the main challenges of using DNN in DH research and the proposed decision model for handling these challenges, and the potential adoption of DNN methods, this paper argues that DH/LIS researchers should expand their arsenal of computational skills and methods. A DH expert must acquire in-depth knowledge in mathematics, software programming and have a deep understanding of the usage of deep neural network frameworks. Therefore, we encourage DH/LIS academic departments to introduce the following topics into their academic syllabus, at the applied (rather than theoretical) level:

- Multivariable calculus (partial derivatives, gradients, high order derivatives),
- Linear algebra (vector space, matrices operations, matrices decompositions),
- Probability (distribution, entropy),
- Statistics (bayesian, parameter estimation, overfitting, and underfitting),
- Mathematical optimization (gradient descent, stochastic gradient descent),
- Unsupervised machine learning (k-means, hierarchical clustering, local outlier factor),
- Supervised machine learning (SVM, logistic regression, naïve bayes, knn),
- Unsupervised and self-supervised deep learning (autoencoders, deep belief networks, generative adversarial networks, embeddings),
- Supervised deep learning (feed-forward, RNN, Self-Attention Network (SAN), CNN),
- Python / R programming (working with data, visualization, ML and DL frameworks, working with GPUs).

Adding these topics to the academic syllabus of DH/LIS experts does not mean that DH/LIS experts will become Computer Science experts, but rather they will be able to comprehend and adapt DL algorithms for their needs. Using this knowledge, DH/LIS experts will no longer be limited to "off the shelf" tools developed for generic open-domain tasks, and will be able to utilize the full potential of the DL algorithms.

Finally, in addition to raising awareness of digital humanities researchers of deep neural networks as the state-of-the-art text analysis method, researchers should be encouraged to generate and release public DH/LIS corpora for training deep neural networks.

REFERENCES:

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16) (pp. 265-283).

Aker, A., El-Haj, M., Albakour, M.D. and Kruschwitz, U. (2012), "Assessing Crowdsourcing Quality through Objective Tasks", in LREC , pp. 1456-1461.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET) (pp. 1-6). IEEE.

Alias-i, L. (2008). 4.1. 0. URL http://alias-i. com/lingpipe.

Almeida, T., Hidalgo, J. M. G., & Silva, T. P. (2013). Towards sms spam filtering: Results under a new dataset. International Journal of Information Security Science, 2(1), 1-18.

Anthony, L. F. W., Kanding, B., & Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. arXiv preprint arXiv:2007.03051.

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015, January). Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015.

Banar, N., Lasaracina, K., Daelemans, W., & Kestemont, M. (2020). Transfer Learning for Digital Heritage Collections: Comparing Neural Machine Translation at the Subword-level and Character-level. In ICAART (1) (pp. 522-529).

Beaudoin, J., & Buchanan, S. (2012). Digital humanities and information visualization: Innovation and integration. Bulletin of the American Society for Information Science and Technology, 38(4), 14-15.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In Neural networks: Tricks of the trade (pp. 437-478). Springer, Berlin, Heidelberg.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). Deep learning (Vol. 1). Massachusetts, USA:: MIT press.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., ... & Bengio, Y. (2010, June). Theano: a CPU and GPU math expression compiler. In Proceedings of the Python for scientific computing conference (SciPy) (Vol. 4, No. 3, pp. 1-7).

Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D. and Panovich, K. (2010), "October. Soylent: a word processor with a crowd inside" In Proceedings of the 23nd annual ACM symposium on User interface software and technology. ACM. pp. 313-322.

Berry, D. M. (2011). The computational turn: Thinking about the Digital Humanities. Culture Machine, 12, 1–22.

Biemann, C., Crane, G. R., Fellbaum, C. D., & Mehler, A. (2014). Computational humanities-bridging the gap between computer science and digital humanities (Dagstuhl Seminar 14301). In Dagstuhl reports (Vol. 4, No. 7). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Blanke, T., Bryant, M., & Hedges, M. (2020). Understanding memories of the Holocaust—A new approach to neural networks in the digital humanities. Digital Scholarship in the Humanities, 35(1), 17-33.

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., & Bjerva, J. (2017). The groningen meaning bank. In Handbook of linguistic annotation (pp. 463-496). Springer, Dordrecht.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

Bunescu, R., & Mooney, R. (2007, June). Learning to extract relations from the web using minimal supervision. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 576-583).

Burgoyne, J. A., Fujinaga, I., & Downie, J. S. (2015). Music information retrieval. A new companion to digital humanities, 213-228.

Chang, S., Zhang, Y., Han, W., Yu, M., Guo, X., Tan, W., ... & Huang, T. S. (2017). Dilated Recurrent Neural Networks. Advances in Neural Information Processing Systems, 30, 77-87.

Chen, C. M., & Chang, C. (2019). A Chinese ancient book digital humanities research platform to support digital humanities research. The Electronic Library, 37(2), 314-336.

Chiron, G., Doucet, A., Coustaty, M., & Moreux, J. P. (2017, November). Icdar2017 competition on post-ocr text correction. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 1423-1428). IEEE.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, October). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1724-1734). Chowdhary, K. R. (2020). Natural language processing. In Fundamentals of Artificial Intelligence (pp. 603-649). Springer, New Delhi.

Cilia, N. D., De Stefano, C., Fontanella, F., Marrocco, C., Molinara, M., & Freca, A. S. D. (2020). An Experimental Comparison between Deep Learning and Classical Machine Learning Approaches for Writer Identification in Medieval Documents. Journal of Imaging, 6(9), 89.

Ciobanu, A. M., Dinu, L. P., Şulea, O. M., Dinu, A., & Niculae, V. (2013, September). Temporal text classification for romanian novels set in the past. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013 (pp. 136-140).

Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167).

Davis, P. T., Elson, D. K., & Klavans, J. L. (2003, May). Methods for precise named entity matching in digital collections. In 2003 Joint Conference on Digital Libraries, 2003. Proceedings. (pp. 125-127). IEEE.

De Smet, H. (2005). A corpus of Late Modern English texts. Icame Journal, 29(29), 69-82.

Deng, L. & Liu, Y. (Eds.). (2018). Deep learning in natural language processing. Springer.

Dereza, O. (2018, October). Lemmatization for Ancient Languages: Rules or Neural Networks?. In

Conference on Artificial Intelligence and Natural Language (pp. 35-47). Springer, Cham.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
Downs, J.S., Holbrook, M.B., Sheng, S. and Cranor, L.F. (2010), "' Are your participants gaming the system?: screening mechanical turk workers". In Proceedings of the SIGCHI conference on human factors in computing systems, ACM. pp. 2399-2402.
Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research, 12(7).
Drucker, J., Kim, D., Salehian, I., Bushong, A. (2014). Introduction to the Digital Humanities. Concepts, Methods and Tutorials for Students and Instructors. Los Angeles: University of California Los Angeles. Available at http://dh101.humanities.ucla.edu.
Eiteljorg, H. (2004). Computing for archaeologists. A Companion to Digital Humanities, 20-30.
Elmalech, A., & Dishi, Y. (2021). Cost-Effective Method for Generating Training Data to Machine Learning Models, iConference 2021 Proceedings.
Elmalech, A., & Grosz, B. (2017). " But you Promised": Methods to Improve Crowd Engagement In Non-Ground Truth Tasks. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (Vol. 5, No. 1).
Elson, D. K., Dames, N., & McKeown, K. R. (2010). Extracting social networks from literary fiction. In Proceedings of the 48th annual meeting of the Association for Computational linguistics, ACL'10 (pp. 138–147). Stroudsburg, PA: Association for Computational Linguistics.
Ess, C. (2004). Revolution? What Revolution? Successes and Limits of Computing Technologies in Philosophy and Religion. Companion to Digital Humanities. Blackwell Companions to Literature and Culture. Malden, MA and Oxford: Blackwell Publishing, 132-144.
Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In Automated Machine Learning (pp. 3-33). Springer, Cham.
Finegold, M., Otis, J., Shalizi, C., Shore, D., Wang, L., & Warren, C. (2016). Six degrees of Francis Bacon: a statistical method for reconstructing large historical social networks. Digital Humanities Quarterly, 10(3).
Finkel, J. R., Grenager, T., & Manning, C. D. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05) (pp. 363-370).
Forte, M. (2015). Cyberarchaeology: a Post-Virtual Perspective. Humanities and the Digital. A Visioning Statement. MIT Press. Boston.
Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.
Glazkova, A. (2020). A Comparison of Synthetic Oversampling Methods for Multi-class Text Classification. arXiv preprint arXiv:2008.04636.
Gold, M. K. (Ed.). (2012). Debates in the digital humanities. U of Minnesota Press.
Gold, M. K., & Klein, L. F. (Eds.). (2016). Debates in the Digital Humanities 2016. U of Minnesota Press.

Gulli, A., & Pal, S. (2017). Deep learning with Keras. Packt Publishing Ltd.

Habernal, I., & Gurevych, I. (2017). Argumentation mining in user-generated web discourse. Computational Linguistics, 43(1), 125-179.

Hajic, J. (2004). Linguistics meets exact sciences (Vol. 42, p. 79). Wiley-Blackwell.

Hall, M(2020). Opportunities and Risks in Digital Humanities Research. In: Carius, Hendrikje; Prell, Martin and Smolarski, René eds. Kooperationen in den digitalen Geisteswissenschaften gestalten. Vandenhoeck & Ruprecht GmbH & Co. KG, Göttingen, pp. 47–66.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hellrich, J., & Hahn, U. (2016, December). Bad company—neighborhoods in neural embedding spaces considered harmful. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 2785-2796).

Hinrichs, E., Hinrichs, M., Kübler, S., & Trippel, T. (2019). Language technology for digital humanities: introduction to the special issue.

Hochreiter, S., & Schmidhuber, J. J. (1997). Long short-term memory. Neural Computation, 9(8), 1–32. https://doi.org/10.1162/neco.1997.9.8.1735

Hutchings, T. (2015), "Digital humanities and the study of religion", in Svensson, P. and Goldberg, D.T.(Eds), Between Humanities and the Digital, MIT Press, Cambridge, MA, pp. 283-294.

Hsueh, G. (2020). Carbon Footprint of Machine Learning Algorithms. Senior Projects Spring 2020. https://digitalcommons.bard.edu/senproj_s2020/296

Jin, J., Yan, Z., Fu, K., Jiang, N., & Zhang, C. (2016). Neural network architecture optimization through submodularity and supermodularity. stat, 1050, 1.

Kingma, D. P., & Ba, J. (2015, January). Adam: A Method for Stochastic Optimization. In ICLR (Poster).

Kittur, A., Smus, B., Khamkar, S. and Kraut, R.E. (2011), "Crowdforge: Crowdsourcing complex work", In Proceedings of the 24th annual ACM symposium on User interface software and technology, ACM, pp. 43-52.

Kollatz, T. (2019). 18 EPIDAT–Research Platform for Jewish Epigraphy. In Crossing Experiences in Digital Epigraphy (pp. 231-239). De Gruyter Open Poland.

Koltay, T. (2016), "Library and information science and the digital humanities: perceived and real strengths and weaknesses", Journal of Documentation, Vol. 72 No. 4.

Krapivin, M., Autaeu, A., & Marchese, M. (2009). Large dataset for keyphrases extraction. University of Trento.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

Kuhn, J. (2019). Computational text analysis within the Humanities: How to combine working practices from the contributing fields?. Language Resources and Evaluation, 53(4), 565-602.

Kumar, A., & Lehal, G. S. (2016). Automatic text correction for Devanagari OCR. Indian Journal of Science and Technology, 9(45).

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for

segmenting and labeling sequence data.
Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.
Li, J., Sun, A., Han, J., & Li, C. (2020). A Survey on Deep Learning for Named Entity Recognition. IEEE Computer Architecture Letters, (01), 1-1.
Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
Liu, Z., Lv, X., Liu, K., & Shi, S. (2010, March). Study on SVM compared with the other text classification methods. In 2010 Second international workshop on education technology and computer science (Vol. 1, pp. 219-222). IEEE.
Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 142-150).
McCarty, W. (2013, July). Becoming interdisciplinary. In DH (pp. 293-295).
McGillivray, B., Poibeau, T., & Ruiz Fabo, P. (2020). Digital Humanities and Natural Language Processing:“Je t’aime... Moi non plus”.
Melis, G., Dyer, C., & Blunsom, P. (2017). On the state of the art of evaluation in neural language models. arXiv preprint arXiv:1707.05589.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009, August). Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (pp. 1003-1011).
Montgomery, D. C. (2017). Design and analysis of experiments. John wiley & sons.
Moreno-Ortiz, A. (2017, April). Lingmotif: Sentiment analysis for the digital humanities. In Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics (pp. 73-76).
Needleman, S. B., & Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. Journal of Molecular Biology, 48(3), 443–453.
Neudecker, C. (2016, May). An open corpus for named entity recognition in historic newspapers. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 4348-4352).
Niculae, V., Zampieri, M., Dinu, L. P., & Ciobanu, A. M. (2014, April). Temporal text ranking and automatic dating of texts. In Proceedings of the 14th Conference of the European Chapter of the

Association for Computational Linguistics, volume 2: Short Papers (pp. 17-21).
Pantel, P., & Pennacchiotti, M. (2006, July). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp. 113-120).
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Desmaison, A. (2019). Pytorch: An imperative style, high-performance deep learning library. In Advances in neural information processing systems (pp. 8026-8037).
Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
Poole, A. H. (2017). The conceptual ecology of digital humanities. Journal of Documentation.
Posner, M. (2013). No half measures: Overcoming common challenges to doing digital humanities in the library. Journal of Library Administration, 53(1), 43-52.
Prebor, G., Zhitomirsky-Geffet, M., Buchel, O., & Bouhnik, D. (2018). A New Methodology for Error Detection and Data Completion in a Large Historical Catalogue Based on an Event Ontology and Network Analysis. Digital Humanities 2018: Book of Abstracts/Libro de resúmenes.
Prebor, G., Zhitomirsky-Geffet, M., & Miller, Y. (2020a). A new analytic framework for prediction of migration patterns and locations of historical manuscripts based on their script types. Digital Scholarship in the Humanities, 35(2), 441-458.
Prebor, G., Zhitomirsky-Geffet, M., & Miller, Y. (2020b). A Multi-dimensional Ontology-based Analysis of the Censorship of Hebrew Manuscripts. *Digital Humanities Quarterly, 14(1*).
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.
Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822.
Ramsay, S. (2016). Who's in and who's out. In Defining digital humanities (pp. 255-258). Routledge.
Rigaud, C., Doucet, A., Coustaty, M., & Moreux, J. P. (2019, September). ICDAR 2019 competition on post-OCR text correction. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 1588-1593). IEEE.
Robinson, L., Priego, E., & Bawden, D. (2015). Library and information science and digital humanities: two disciplines, joint future?. Re-inventing information science in the networked society.
Rokach, L. and Maimon, O. (2015) *Data mining with decision trees: theory and applications*. 2nd editio. Hackensack New Jersey: World Scientific.
Rommel, T. (2004). Literary studies. In A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell. http://www.digitalhumanities.org/companion/
Rossum, G. (1995). Python reference manual.
Rubinstein, A. (2019). Historical corpora meet the digital humanities: the Jerusalem Corpus of

Emergent Modern Hebrew. Language Resources and Evaluation, 53(4), 807-835.
Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323, 533–536. 159, 330
Russell, I. G. (2011). The role of libraries in Digital Humanities. Erişim adresi: http://www.ifla.org/past-wlic/2011/104-russell-en. pdf.
Saltz, D. Z. (2004). Performing arts. A Companion to Digital Humanities, 121-131.
Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. IEEE Transactions on Neural Networks, 20(1), 61-80.
Schulz, S., & Ketschik, N. (2019). From 0 to 10 million annotated words: part-of-speech tagging for Middle High German. Language Resources and Evaluation, 53(4), 837-863.
Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673–2681. https://doi.org/10.1109/78.650093
Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24(4), 35-43.
Skadiņš, R., Tiedemann, J., Rozis, R., & Deksne, D. (2014, May). Billions of parallel words for free: Building and using the eu bookshop corpus. In Proceedings of LREC.
Snow, R., O'Connor, B., Jurafsky, D. and Ng, A.Y. (2008), "Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks", In Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp. 254-263.
So, D. R., Liang, C., & Le, Q. V. (2019). The evolved transformer. arXiv preprint arXiv:1901.11117.
Sorokin, A. and Forsyth, D. (2008), "Utility data annotation with amazon mechanical turk". In 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp. 1-8.
Spinaci, Gianmarco, Colavizza, Giovanni, & Peroni. (2019). List of Digital Humanities journals (Version 0.1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3406564.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15, 1929–1958. https://doi.org/10.1214/12-AOS1000
Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.
Suissa, O., Elmalech, A., & Zhitomirsky-Geffet, M. (2019). Toward the optimized crowdsourcing strategy for OCR post-correction. Aslib Journal of Information Management, *72(2),* 179-197.
Suissa, O., Elmalech, A., & Zhitomirsky-Geffet, M. (2020). Optimizing the neural network training for OCR error correction of historical Hebrew texts. iConference 2020 Proceedings.
Sula, C. A. (2012). Visualizing social connections in the humanities: Beyond bibliometrics. Bulletin of the American Society for Information Science and Technology, 38(4), 31-35.
Sula, C.A. (2013), "Digital humanities and libraries: a conceptual model", Journal of Library Administration, Vol. 53 No. 1, pp. 10-26.
Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. International Journal of Computer Vision.

Svensson, P. (2010). The landscape of digital humanities. Digital Humanities, 4(1).
Tanasescu, C., Kesarwani, V., & Inkpen, D. (2018, May). Metaphor detection by deep learning and the place of poetic metaphor in digital humanities. In The Thirty-First International Flairs Conference.
Thomas, W. G. (2004). Computing and the historical imagination. A companion to digital humanities, 56-68.
Thompson, P. (2017). The voice of the past: Oral history. Oxford university press.
Thyaharajan, S. K., Sampath, K., Durairaj, T., & Krishnamoorthy, R. SSN NLP at CheckThat! 2020: Tweet Check Worthiness Using Transformers, Convolutional Neural Networks and Support Vector Machines.
Todorov, K., & Colavizza, G. (2020, January). Transfer Learning for Named Entity Recognition in Historical Corpora. In CLEF (Working Notes).
Torrey, L., & Shavlik, J. (2010). Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques (pp. 242-264). IGI global.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
Vijayan, V. K., Bindu, K. R., & Parameswaran, L. (2017, September). A comprehensive study of text classification algorithms. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1109-1113). IEEE.
Wang, Q., Luo, T., Wang, D., & Xing, C. (2016, July). Chinese song iambics generation with neural attention-based model. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (pp. 2943-2949).
Warwick, C. (2012). Institutional models for digital humanities. Digital humanities in practice, 193-216.
Williams, R. R., Hunt, S. C., Heiss, G., Province, M. A., Bensen, J. T., Higgins, M., ... & Hopkins, P. N. (2001). Usefulness of cardiovascular family history data for population-based preventive medicine and medical research (the Health Family Tree Study and the NHLBI Family Heart Study). The American journal of cardiology, 87(2), 129-135.
Won, M., Murrieta-Flores, P., & Martins, B. (2018). ensemble named entity recognition (ner): evaluating ner Tools in the identification of Place names in historical corpora. Frontiers in Digital Humanities, 5, 2.
Wooldridge, R. (2004). Lexicography. A Companion to Digital Humanities, 69-78.
Yang, E., Ravikumar, P. K., Allen, G. I., & Liu, Z. (2013). On Poisson graphical models. Advances in Neural Information Processing Systems, 26, 1718-1726.
Yang, S., Wang, Y., & Chu, X. (2020). A Survey of Deep Learning Techniques for Neural Machine Translation. arXiv preprint arXiv:2002.07526.
Yang, Y., Bansal, N., Dakka, W., Ipeirotis, P., Koudas, N. and Papadias, D. (2009), "Query by document", In Proceedings of the Second ACM International Conference on Web Search and Data Mining, ACM, pp. 34-43.
Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237.
Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural

language processing. ieee Computational intelligenCe magazine, 13(3), 55-75.
Zaagsma, G. (2013). On digital history. BMGN-Low Countries Historical Review, 128(4), 3-29.
Zampieri, M., & Becker, M. (2013). Colonia: Corpus of historical portuguese. ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research, 5, 69-76.
Zhitomirsky-Geffet, M., & Prebor, G. (2019). SageBook: Toward a cross-generational social network for the Jewish sages' prosopography. Digital Scholarship in the Humanities, 34(3), 676-695.
Zhitomirsky-Geffet, M., Prebor, G. & Miller, I. (2020). Ontology-based analysis of the large collection of historical Hebrew manuscripts. Digital Scholarship in the Humanities, 35(3), 688-719.
Zhou, G., & Su, J. (2002, July). Named entity recognition using an HMM-based chunk tagger. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 473-480).

WORKSHOP: LINKED DATA FOR DIGITAL HUMANITIES

Annemieke Romein
Post-doctoral researcher, Ghent University (Belgium)/ KB National Library of the Netherlands

I am immensely grateful for participating in the 2019 version of the Digital Humanities Summer School. I learned a great deal on Digital Humanities over the past year and I am still in awe at the wide range of digital applications that are developed to be applied in the Humanities. These make research more accessible, visible and understandable. As a post-doctoral researcher (ECR) in early modern legal/ political-institutional history, I was 'raised' with traditional methodology. I am currently working on a project at Ghent University, which (manually) categorises the early modern legislation from the province of Flanders and Holland. At the KB National Library of the Netherlands, I am working on a way to apply machine-learning to this categorisation of legislation. Since I use the same categorisation as the Max-Planck-Institute für europäische Rechtsgeschichte (Legal History) it makes sense to look for means to ensure the use of same definitions within our data sets.

Thus, I attended the workshop Linked Data for Digital Humanities (LD4DH), run by Dr Terhi Nurmikko-Fuller. She is an amazing teacher, with lots of humour (not just in wearing sparkling clothes while discussing SPARQL) and a tremendous amount of experience. She provided us with a step-by-step theoretical background and a lot of chances to get dirty with hands-on experiences. That was not always easy, as Open Source programs make you have a fit – but together with the other teachers, she takes the time to solve our malfunctioning computers.

On Monday, we kicked off with learning the difference between a relational database and Linked (Open) Data, triples,RDF and ontologies. This was somewhat challenging as it feels like learning a whole new language (while English is not my native tongue anyway). We finished our first day with discussing Linked Data in numismatics. It was really useful to have such a practical talk at the end of the day, as it gave some insights into the applicability. Tuesday morning was thought-provoking as we had to write and discuss an ontology for a dataset we were provided with. It was very challenging to think about the data on a meta-level and distinguish between Classes and properties, we had to put this into a tool called Protege (and clean out the given data through OpenRefine). Later on, we learned how this could be applied within Musicology and that entire music pieces can be written using Linked Data. Wednesday had its challenges as many of us struggled to run WebKarma. The program is really useful, as soon as you have your computer run it. When the data is ready, you need to export it to Blazegraph – another interesting program in which you can query your data. I really admire the patience of Terhi, John, and Graham in troubleshooting with many of us, but in the end, managing to solve about every single problem we encountered and – while at it – teach us the use of Web Karma and Blazegraph too! As soon as you have the hang of it, a world of opportunities opens up. An illustration of such an opportunity was given that day by Stephen Downie, with the HathiTrust.

Wearing sparkling trousers (on the hottest day ever), Terhi gave us an introduction to SPARQL querying – through which you can query to find whatever data has been linked within your entire data set. Of course, you can also use other programming languages, but for those who do not know, after having worked with all the other tools SPARQL is very functional to start with. Having 'survived' all that, we were presented with many other applications – such as Recogito (to find geographical references and plot these on a map) and Knowledge Graph (linking images from the British Museum). The final lecture dealt with the model of CIDOC-CRM – which helped to understand the logic behind data structuring. This nicely wrapped up the challenge that Linked Data poses.

This workshop is extremely useful and practical for those working with datasets that link to other institutes or datasets that should be versatile in its usage (a lot of data of which you know you will need in various combinations). If you want your computer to just do as you order it, you are in the wrong place because the Open Source tools may need some special attention and care before they work the way you expect.

It was a great week to meet like-minded people from various disciplines, countries, and stages of research. Keble College is a fantastic stage (though our room was 'somewhat hot'), marvellous lunches, and a wide arrange of optional presentations/ discussion fora to attend. A brilliant place to crash-course your knowledge of Digital Humanities in general, and many topics in specific. There is no doubt in my mind that if you want to learn more about Digital Humanities: this is the place to go.

SPATIAL DESIGN OF PROTESTS: EXPLORING THE LITERARY LINEAGE OF PROTEST SITES

Apsara Bala[1] and Nirmala Menon[2]

1: Doctoral Student, Digital Humanities and Publishing Studies Research Group, SHSS, IIT Indore. Email: phd2101161002@iiti.ac.in; 2: Professor and Head, SHSS, IIT Indore. Email: nmenon@iiti.ac.in

ABSTRACT :

The increasing scale of interest to research in the field of protest has addressed the different aspects of the collective action (social, cultural, political, etc.), but little gained attention to the 'space' in which the materialization and manifestation take place. Public spaces as a practical as well as symbolic place connect citizens and have always been central to social movements and political protests. But the growing hostility and colonial attitude towards protests coupled with the rhetoric of disruption have affected the evolvability of the phenomena. In this way, from Tahrir Square to Occupy Wall Street, Tiananmen Square to Shaheen Bagh, and Farmers' protests, accompanying a multidisciplinary approach the research efforts have involved discussion on protest, public space, people, literature, and architecture. The accessibility of the public in the central space (Delhi) that had the opportunity in holding the state power accountable and respond to their demands became obsolete (Guha, 2007). The displacement of the designated space for protest also distanced the visibility of the dissent from those in power. This unacceptance that the central space possesses has laid bare the socio-spatial inequalities between government and citizens. Thus, with the exploration of the spatial nitty-gritty of protest, this research focuses on securing the availability, accessibility, and reachability of public spaces in the inclusion of all the members of society.

The research poses questions on the importance of a particular space for protests and the criticism against the designation of spaces for the same. The aim of the study includes analyzing the spatial aspect of protests and the role of accessible public spaces and governmental interventions in them. As a part of the methodology, the research accommodates textual analysis (literature that has acknowledged/ described/ philosophized the spatial aspect of the protests) while intending to explore all the spatial viewpoints for protests. The simultaneous mapping method helps in visualizing the historical displacement and transformation of accessibility – from Rajpath to Farmers' protest taking place at the edges of the city due to spatial denial (Zuberi, 2020). The paper, therefore, constitutes a two- part analysis – theorizing the public space in relation to protests and the practical part that comes with mapping out public spaces on an urban cityscape. The designation of places for protest, the shrinking of public spaces, and calling upon restrictions can land on monitoring and controlling the designated spaces granting only permissive accessibility (Mantri, 2021). Also, the acceleration of privatized public spaces having the illusion of access has steadily replaced the real public spaces. This study in connecting and analyzing the relation of public spaces with protests, in this way, tries to explore the myriad intricacies growing around the notion.

Public space, literature, people, and architecture can be viewed from a wide range of perspectives, but here protest is the key concept to tie them up in one unified whole. This research becomes pertinent for being devoted to a social cause in talking about the accessibility of public spaces through a discussion on protests and providing scope for future research to work on many unpacked/ unheard stories from the protest sites. Again, the multidisciplinary approach running along with a multilingual analysis of literature has doubled up for the inclusion of multiple perspectives in the study.

KEYWORDS

Occupy Movements, Public space, Accountability and accessibility, Displacement, Protesters, Socio-spatial inequalities, Urban design

Works Cited –

"From Kingsway to Rajpath: How Independent India Made a British Imperial City Its Own." The Indian Express, 17 May 2010, https://indianexpress.com/article/research/from-kingsway-to-rajpath-india-gate-how-independent-india-made-a-british-imperial-city-its-own-7315789/.

Guha, Ramachandra. India After Gandhi: The History of the World's Largest Democracy. Picador India, 2017.

Hatuka, Tali. The Design of Protest: Choreographing Political Demonstrations in Public Space. First edition, University of Texas Press, 2018.

Lefebvre, Henri. The Production of Space. Blackwell, 1991.

Mantri, Mamta. Cities and Protests: Perspectives in Spatial Criticism. Cambridge Scholars Publishing; Unabridged edition, 23 July 2010.

Mehrotra, Neha. "How Jantar Mantar Killed the Spirit of Protest." Https://Www.Outlookindia.Com/, 2 Feb. 2015, https://www.outlookindia.com/website/story/how-jantar-mantar-killed-the-spirit-of-protest/299180.

WHY MAP LITERATURE? GEOSPATIAL PROTOTYPING FOR LITERARY STUDIES AND DIGITAL HUMANITIES

Randa El Khatib, Marcel Schaeben

Final Authors' copy

ABSTRACT:

By focusing on the process of building A Map of Paradise Lost—a geospatial humanities text-to-map project that visualizes the locatable places in John Milton's Paradise Lost— this paper addresses the question "why map literature?" and demonstrates how the process of research prototyping is in itself a form of knowledge production. Through a series of prototyping moments, we address how the different steps involved in building a geospatial humanities project can produce new knowledge about the fields it relates to: literary studies and digital humanities. The prototyping moments make arguments that advance our understanding of Milton's Paradise Lost, approaches to data visualization for cartographic comparison in and beyond DH, and models for interdisciplinary collaboration.

KEYWORDS:

literary mapping, geospatial prototyping, scholarly communication, Paradise Lost Geospatial humanities, literary mapping, research prototyping, scholarly communication, Paradise Lost

Introduction

Geospatial humanities is a significant and rapidly growing branch of the digital humanities and constitutes the practice of applying Geographical Information Systems (GIS) and other quantitative technologies to the study of the representation of spatiality in texts, often to literary or historical content. A multitude of geospatial humanities projects involve geovisualizing literary texts. As a fairly recent research area, the contribution of literary mapping is still being established: why do we map literary texts? (Cooper, Donaldson, and Murrieta-Flores 2016, 9). Scholars have questioned the value of geospatial humanities projects for digital humanities and literary studies, often with inconclusive remarks (Piatti, 2016). This uncertainty is understandable in a nascent field that relies so strongly on digital tools and methods that are as diverse as they are rapidly developing; no wonder why it is difficult to single out their contribution to scholarship in an ever-changing medium. Allison Muri (2016) claims that "…in the digital humanities experimental studies are important, valid, and necessary trials as we test new methods in a still nascent field. We cannot proceed without experiments and testing of hypotheses. We also need to ask—and answer—(to recast Alan Liu's (2013) important question about the Digital Humanities), what is the *meaning* of a literary GIS to literary studies and textual scholarship?" (2016, n.p.). Until we can define what geospatial humanities encompasses and set its boundaries in order to make overarching claims about its scholarly landscape, we ought to concentrate on a case-by-case exploration of the significance of individual projects to the research areas they relate to. This very type of examination itself can help define the field's parameters.

We examine the scholarly contribution of geospatial humanities to both fields that are involved in this particular interdisciplinary instance—digital humanities and literary studies—by looking at the prototyping

process of building a literary map of John Milton's *Paradise Lost*. Although research prototyping may have some overlaps with other established forms of knowledge production, such as developing a critical edition of a work or writing an article or monograph, many of the processes, such as data gathering and interpretation, collaboration, and platform design, demand a different way of approaching the text. At the core, the steps necessary for creating any of the above require engaging with the text directly, as well as with the cultural and historical materials surrounding the text. By recasting Alan Galey and Stan Ruecker's (2010) momentous inquiry into "how a prototype argues to "how a *geospatial* prototype argues," we address how the different steps involved in building a geospatial humanities project can produce new knowledge about the fields it relates to: literary studies and digital humanities. This study is carried out by addressing specific prototyping moments—critical decisions about data gathering and structuring, as well as decisions about the features of the app—that demonstrate how the process of building is in itself a form of knowledge production that can grant new ways of engaging the text and imagining technical solutions to collaboratively visualize complex, multilayered, literary space.

A Map of Paradise Lost

A Map of Paradise Lost is an open access online project that situates John Milton's *Paradise Lost* in its specific historical moment—the seventeenth century—published in a map-oriented culture that was at the peak of the development of a cartographic consciousness in Europe. In "Milton's Maps," Morgan Ng (2013) argues that there is a "tendency among current literary scholars, despite enormous interest in the 'cartographic imagination' in Renaissance writing, largely to ignore the texts' actual visual counterparts. To explain the textual form of *Paradise Lost* requires equally close attention to the images which permeated Milton's mimetic consciousness, even after the onset of his blindness." (2013, 428). Ng (2013) points beyond the textual references that are typically read alongside *Paradise Lost* to actual maps, such as the map of biblical lands found in the paratext of Milton's own family map, a 1612/13 printing of the King James Bible—one with which Milton would have no doubt been familiar with over the course of his life and that, according to Ng (2013), influenced his spatial thinking about the places mentioned in biblical accounts and depicted in *Paradise Lost*.

Paradise Lost creates a rich and complex world that draws on multiple histories, with references to places from classical antiquity, biblical accounts, and Milton's contemporary world, to name a few. Milton's allusions to places are explained in critical notes of some scholarly editions, where editors contextualize and interpret the complex and multilayered references to places and their significance. As a starting point for contextualizing the significance of places mentioned, authors consulted *The Complete Poetry and Essential Prose of John Milton* (2007) edited by William Kerrigan, John Rumrich, and Stephen M. Fallon, and expanded to other editions and critical sources as listed in the bibliography. Still, the superimpositions of multiple allusions upon the terrestrial world of the epic poem is challenging to keep track of, especially when approaching the full work. *A Map of Paradise Lost* is the first project of its kind that grants visual access to the world of "geographical continuity" that permeated Milton's spatial consciousness. This worldview refers to the prevalent notion of historical sequence of a seventeenth-century English audience, namely the conviction that biblical events, like historical ones, progressed on a linear spectrum of geographical continuity, meaning that the land that the Ottoman Empire occupied in the seventeenth century is the same land in which biblical and classical accounts took place

(Ng 2013, 433). In the project, geographical continuity is demarcated by historical maps that are meant to evoke these worlds imagined by Milton.

The project is an application of Geographical Information System (GIS) techniques to the text of *Paradise Lost* that is meant as an exploratory tool for researchers, students, and readers to investigate the complex and multilayered space of the epic poem. The map includes the locatable platial references, or references to places, with explanatory excerpts from editorial notes that describe their significance. Every place name is connected to the passage in *Paradise Lost* in which it appears (see Figure 1), with multiple passages for places that are mentioned more than once. Rather than simply overlaying platial references on a modern map, the project attempts to visualize some of the many temporalities that are merged into the world of *Paradise Lost*, captured in the georectification (matching points on a map image with corresponding points on a map that exists in GIS) of two maps: the "Map of Biblical Lands" from the King James Bible and the map of "The Turkish Empire" by John Speed (1626). The imperfect fit between the historical maps and GIS interface, seen in some stretched out parts of the rectified maps, is a product of the misalignment of underlying map projections, since the Ptolemaic coordinate system does not align neatly with the Mercator projection (Ng 2013, 430).

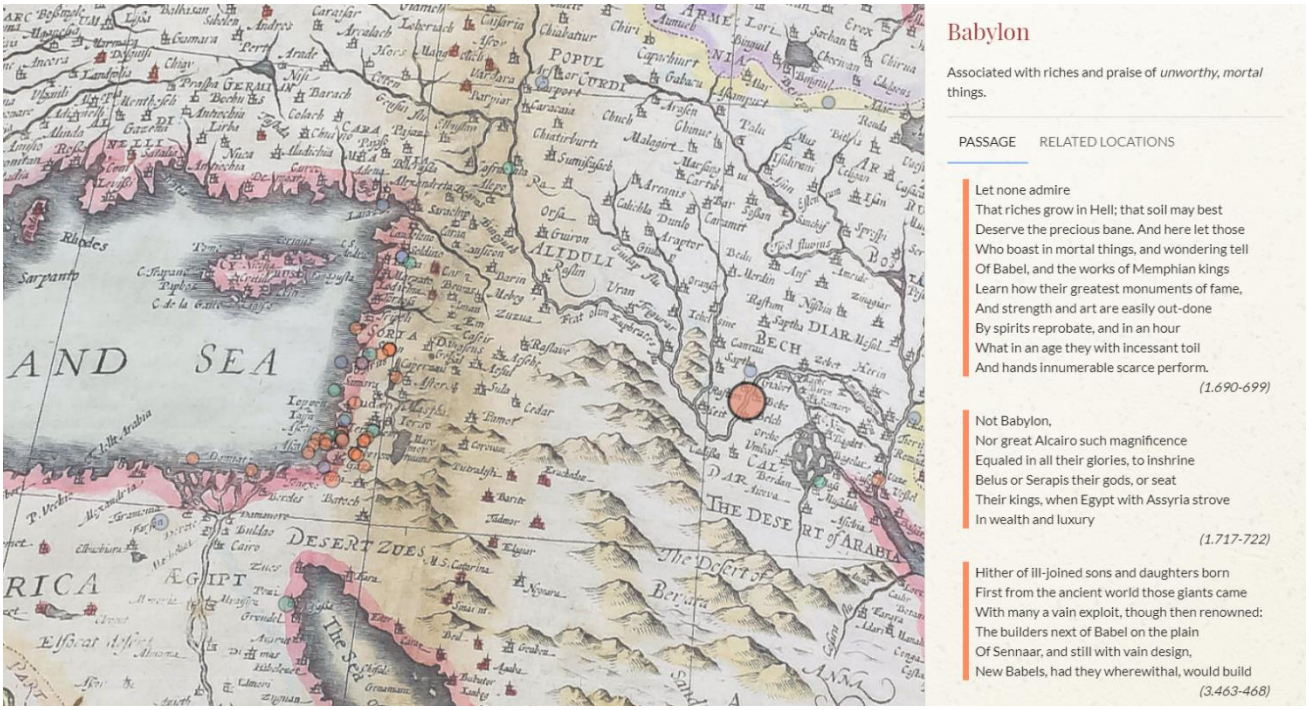


Figure 1

A Map of Paradise Lost: selection of Babylon shows the multiple passages and description of its significance in *Paradise Lost*.

Digital projects reflect the data that fuels them. For example, *The Atlas of Early Printing* (Prickman et al. 2013) traces the development of print in fifteenth-century Europe with respect to other variables such as trade routes, universities, and paper mills by demarcating each variable with its own color on the map alongside a timeline—a simple and effective solution for its purpose—with additional information about the city, year of first printing, name of printer, first work printed, and the *Incunabula Short Title Catalogue* (ISTC) entry. Likewise, Stanford University's *Authorial London* (Evans et al. 2011), a literary geography project that maps

references to London in the works and biographies of notable authors who lived in, travelled to, or wrote about London, includes the title, author, and year of publication of works that refer to places in London. *Authorial London* also features options such as word searches in the corpus or by region, and, like *A Map of Paradise Lost*, includes the passages that refer to those places. Unlike other types of geospatial projects in which references to places alone might be enough to communicate a purpose, such as the movement of wind currents in a region, the humanistic representation of place in literary maps almost always requires contextualization. A point on a map that virtually means "this is Paris" for Paris may not tell us much. In a humanistic context, we often engage with place rather than space. Tim Cresswell defines place as "how we make the world meaningful and the way we experience the world. Place, at a basic level, is space invested with meaning in the context of power. This process of investing space with meaning happens across the globe at all scales and has done throughout human history" (2015, 19). This definition could be extended to literary maps by focusing on the different narratives that give place meaning, of which there are manifold in *Paradise Lost*, where layered allusions work to contextualize the platial references across multiple temporalities at once. For example, Ormus and Ind are grouped together as sites of wealth and associated with Satan's seat in Hell (PL 2.1–5), first assimilated with Islamic Empires, and, by virtue of extension, with Catholic Rome (Quint, 2014; Lim 2010).

Scholarly communication and literary mapping

The digital medium has ushered an age of experimentation with forms of scholarly communication, from research to dissemination. In the digital humanities, this experimentation has taken many forms, including digital editions, online journals and encyclopaedias, dynamic databases, and software prototypes. Digital editions generally provide content in more accessible and networked ways. Hyperlinked information, different media affordances, zooming options, and other features of the digital augment the reading experience. Created for different purposes, digital editions can grant the public and scholars access to rare or brittle manuscripts; advance knowledge and understanding of a work; and experiment with alternative forms of reading and knowledge production. Patrick Sahle (2016) differentiates scholarly digital editions from non-scholarly and print scholarly editions. Some of the primary differences, according to Sahle (2016), are a result of the affordances of the different mediums and the changing scholarly values that they supply; one of the results is that instead of editions claiming their primary purpose to be an authoritative reading and final statement on a subject, the moment of publication of a digital edition is quite fluid since it can be published iteratively rather than finally, in contrast to print editions. Namely, it "becomes a permanent but potentially always changing documentation of an ongoing examination and processing of the objects in question. In this way, the edition as a publication is a process rather than a product." [emphasis added] (n.p.). An example of a scholarly digital edition is *The Grub Street Project*, a social edition of eighteenth-century London that, through the mapping of literature, trade, and print culture of the time, aims to "create both a historically accurate visualization of the city's commerce and communications, and a record of how its authors and artists portrayed it." (*The Grub Street Project,* Home Page). *The Grub Street Project*, like many other digital humanities projects, has continued expanding over the years, with more recent additions, such as a new interface for the maps and some new editions. This iterative element is also true for other types of digital projects; for example, *Authorial London* is

planning on releasing its underlying infrastructure, *Authorial {x}*, in open source —not for adding new editions like *The Grub Street Project*—but rather to allow users to create similar projects for other cities.

Digital projects can blur the boundaries that separate different scholarly genres and produce hybrids that amalgamate different elements. Thinking about the distinction that Sahle (2016) makes between scholarly digital editions as final words in their fields and non-scholarly digital editions that encourage creativity and iterative development, *A Map of Paradise Lost* is essentially a text to map project that visualizes and interprets the spatiality of *Paradise Lost*. In doing so, we ask ourselves whether this is a step towards what we may call a geo-edition, namely, a thematic geo-edition that, instead of taking the text as a whole, visually reconstructs the spatiality of the text—in this case, *Paradise Lost*—by focusing on spatial data, close reading, and editorial context. The process of building the project relies on gathering the data within specific parameters while consulting the primary text and secondary works on the one hand, and thinking critically about questions concerning access and readability through its functionalities, design, and visualizations on the other. In all stages of this collaboration, we were considering what the prototyping process yields in terms of scholarly knowledge, building on the notion of prototyping as a way of thinking, not unlike the well-established and accepted practice of writing as a way of thinking (Ruecker 2015, 3). The resulting project is a hybrid that draws on some editorial practices, such as transmediating the text, framing it, and offering critical insight; however, instead of providing a diplomatic transcription of *Paradise Lost*, it draws on content that is specifically related to platial references and is meant to facilitate a deeper understanding of the spatiality of *Paradise Lost*.

How a geospatial prototype argues

A Map of Paradise Lost is both a prototype and an app; this is not to say that it is not in a fully developed state in its current form, but rather that it has an in-built capacity to continue expanding and to support additional layers that offer different interpretations of *Paradise Lost*'s spatiality. We framed the building process in accordance with Alan Galey and Stan Ruecker's approach by questioning, from the beginning, "how the process of designing may be used simultaneously for creating an artifact and as a process of critical interpretation, and whether new form of digital objects, such as interface components and visualization tools, contain arguments that advance knowledge about the world" (2010, 406). The map itself is the result of the critical inquiries that shaped it, centered around three main prototyping moments that seek to advance our understanding of Milton's *Paradise Lost*, approaches to data visualization for cartographic comparison in and beyond DH, and models for interdisciplinary collaboration, respectively corresponding to arguments: 1) that Milton often imposes moral categories onto place names at use, and that this framework, arrived at through close reading, can be visualized and expanded, 2) that parallel, synchronized maps offer a more intuitive and productive way to compare across maps, broad temporal categories, and data layers, and 3) that platforms can be prototyped to enhance interdisciplinary collaborative practices by minimizing labour for humanists and developers, and maximizing productivity and accessibility. In all three cases, the prototyping moments demonstrated here can be reproduced in other contexts and for other projects.

Let us return to our borrowed question at the outset: "Why map?" (Cooper et al. 2016, 9). As Sebastién Caquard (2011) points out, "Neither cartography nor narrative on their own can capture the essence of place: both are required to get a better sense of it" (224). Especially for works that rely on some understanding of spatial movement and change over time, there is an interdependence between narratives and cartography in

representing place; together, they can provide a more thorough understanding. According to Barbara Piatti and Lorenz Hurni (2011, 222), "Through literary geography, we learn more about the production of places, their historical layers, their meanings, functions and symbolic values. If places emerge from a combination of real elements and fictional accounts, then literary geography and literary cartography can work as a very effective eyeopener." Building on Cooper and Gregory's (2011, 90) take, we "use GIS technology as a tool for critical interpretation rather than mere spatial visualization." Together, the project is meant to offer visual access to the different worlds on the imagined surface of the Earth in *Paradise Lost* in order to provide a contextualization of the historical and biblical framing of the work, and beyond that to serve as an exploratory tool for users to generate their own questions about the spatiality of the epic poem.

During the course of development, it has become a general approach in our design decisions to include many variables in visualizing the data and give the user a choice to switch between different versions or to turn them off altogether. According to Luchetta (2018), the option to switch between project features leads to an experimental, iterative development process that could pose new questions about the visualization of multivariable spatial data and might lead to some new insights. But above all, it also puts users in the position to use the tool in ways and for purposes that we have not envisioned yet.

Prototyping moment 1: Close reading meets map visualization

A generic distinction of sixteenth- and seventeenth-century works that describe places, such as travelogues or chorographies, is their insistent contextualization of place names at use. Scholars have traced many works from which Milton draws spatial references and the ideological connotations with which they are associated. Some main references are from Peter Heylyn's (1657) *Cosmographie in Four Books*, George Sandys' *Relation of a Journey* (1610) and Thomas Fuller's *Pisgah-Sight of Palestine* (1662). At the same time, Grant McColley (1937) demonstrates how Milton consulted one of the most widely-read books of his time, the *Pansebeia: or, A View of all Religions in the World* by Alexander Ross, when writing the demons and their followers in Book I. McColley (1937) identifies how both works follow the pattern of the epic catalogue, in which "the heathen deities are named, their characteristics identified, the places of their worship given, and the practices of their followers described and condemned" (181). The similarities between the language, when placed side-by-side, is striking. One of many examples provided by McColley (183) is when Milton describes Dagon:

Dagon his name, sea-monster, upward Man
And downward Fish: yet had his Temple high
Rear'd in Azotus, dreaded through the Coast
Of Palestine, in Gath and Ascalon,
And Accaron and Gaza's frontier bounds
—(Paradise Lost 1.462–466).
Dagon from Dag a Fish, because
from the navel downward he
was made in the form of a fish,
but upward like a man this
was a great idol among the

Philistines

—(Pansebeia 66; McColley *183).*

A glaring reality of the seventeenth century is that places of biblical accounts and classical antiquity were under Islamic rule, primarily occupied by the Ottoman Empire, at that time a power that posed the biggest threat to Christian England and Europe. Drawing a connection between Satan and an Ottoman sultan is Walter Lim, who cites descriptions of Satan as the "mighty Chief" (PL 10.455) of his "great consulting Peers" (10.456) and compares this group to the "dark Divan" (10.457), a reference that evokes the Turkish council of state (213). As Lim continues: "By portraying hell as the infernal archetype of the Turkish political economy, Milton demonizes the Ottoman Empire and the Muslim faith that it practices, defends, and seeks to disseminate by the point of the sword" (213). However, the fear of the Ottoman Empire lay beyond just military power and religious difference, but in the fear of actual conversion. As Campbell puts it:

Muslims were people that Christians could become; good angels, as in Paradise Lost, could fall into apostasy and become bad angels, just as Christian captives could become Muslims. This was phrased not in terms of "embracing Islam" as it would be now, but rather in terms of "turning Turk." The phrase is telling, because it defers not to religious authority but to the dominant political power. Christians converted because Islam was stronger. (2007, 18)

The anxiety of conversion was a well-founded fear echoed by Daniel Viktus in *Turning Turk* (2003), where he insists that English anxiety about the Ottoman Empire was pushed even further by witnessing the inclusive social system that was adopted by the Turks that had resulted in mass conversion of English Protestants to Islam, and reached its peak when the Ottoman Empire started exponentially advancing its territories towards Hungary, Poland and even Germany (17). By extension, this anxiety of conversion is echoed in much of literature of the early modern period, including in *Paradise Lost*, where, as Campbell puts it, Milton's epic "focuses on the horror of angels and our first parents converting from innocence to guilt, in effect enacting the spiritual equivalent of turning Turk" (19). It is convenient then, that notable landmarks of the Ottoman Empire are so directly linked to pagan worship; places that, in *Paradise Lost*, symbolize the postlapsarian world and the straying away from the "true" path.

By visualizing the spatiality of *Paradise Lost*, Milton's geographical critique becomes more apparent. Not only do places carry a moral connotation, but they are also often grouped together under the same moral category. By extracting data from *Paradise Lost* with attention to moral collocations with place name references, their significance, and how they are grouped together, our first prototyping moment geovisualizes a close reading of *Paradise Lost*. Grouped into broad categories under negative, positive, and neutral, corresponding to orange, green, and blue, these colors are meant as an invitation to the epic poem's more complex context through the passages and editorial comment (see a more detailed explanation of this process in "Mapping the Moralized Geography of *Paradise Lost*" by El Khatib and Currell, 2018). One thing that becomes apparent, even at a glance, is that most places are, in fact, collocated with moralizing content. For example, Rabba, Argob, Basan, and Arnon (see Figure 2) are all grouped together in the same passage, described in Book I as places of pagan worship, followers of Moloch:

First, Moloch, horrid King, besmeared with blood
Of human sacrifice, and parents' tears;

Though, for the noise of drums and timbrels loud,
Their children's cries unheard that passed through fire
To his grim idol. Him the Ammonite
Worshiped in Rabba and her watery plain,
In Argob and in Basan, to the stream
Of utmost Arnon. Nor content with such
Audacious neighbourhood, the wisest heart
Of Solomon he led by fraud to build
His temple right against the temple of God
On that opprobrious hill, and made his grove
The pleasant valley of Hinnom, Tophet thence
And black Gehenna called, the type of Hell.
—(Paradise Lost, I. 391–405).

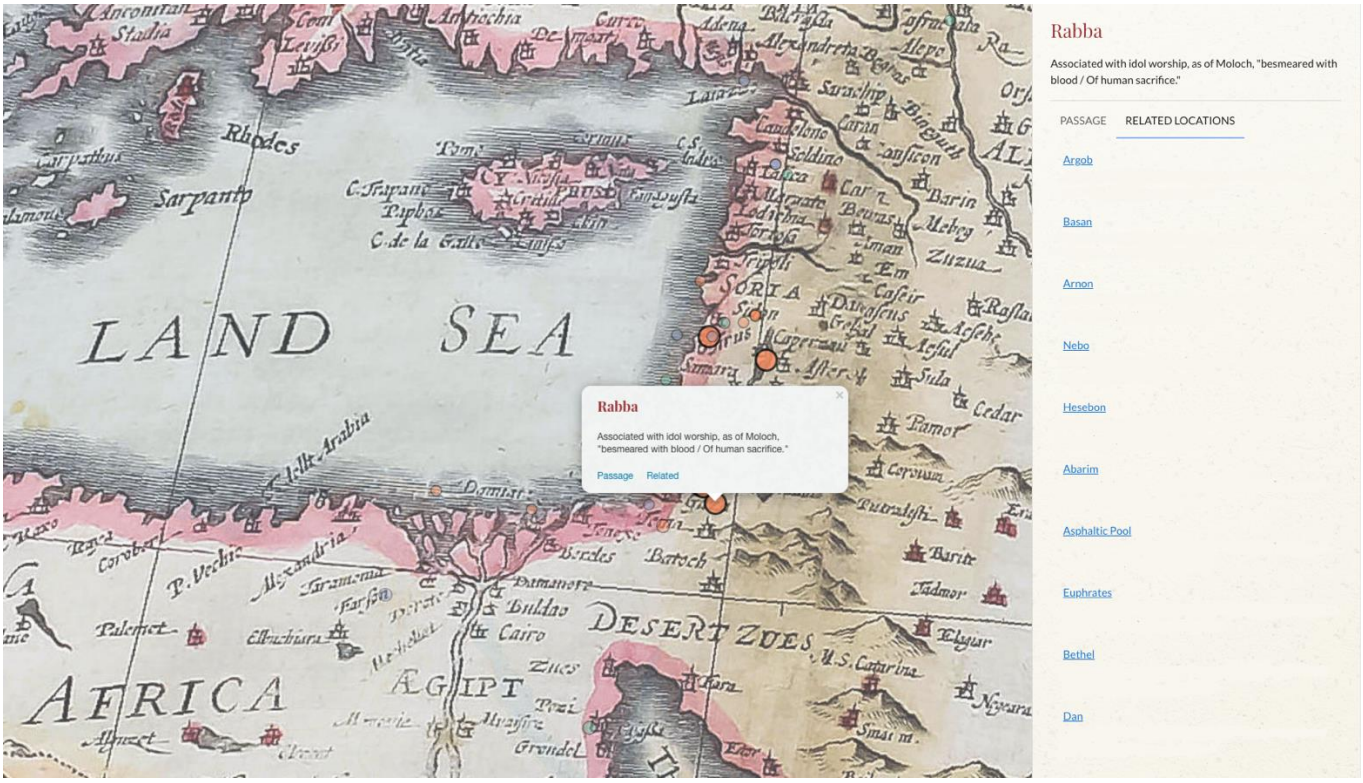


Figure 2

A Map of Paradise Lost: spatial grouping of related places (in this case Rabba with places associated with idol worship).

The spatial grouping itself happens across all places related to the broader category, and therefore references all places associated with idol worship and pagan cults, like Nebo, Hesebon, Abarim, the Asphaltic Pool (grouped together as worshipers of Chemos) and Euphrates, Bethel, and Dan (collocated with the Hebrews' worship of the golden calf). In cases where the place appears more than once in the text of the epic poem, each mention is treated separately and has a passage extract and moral collocation (Figure 3).

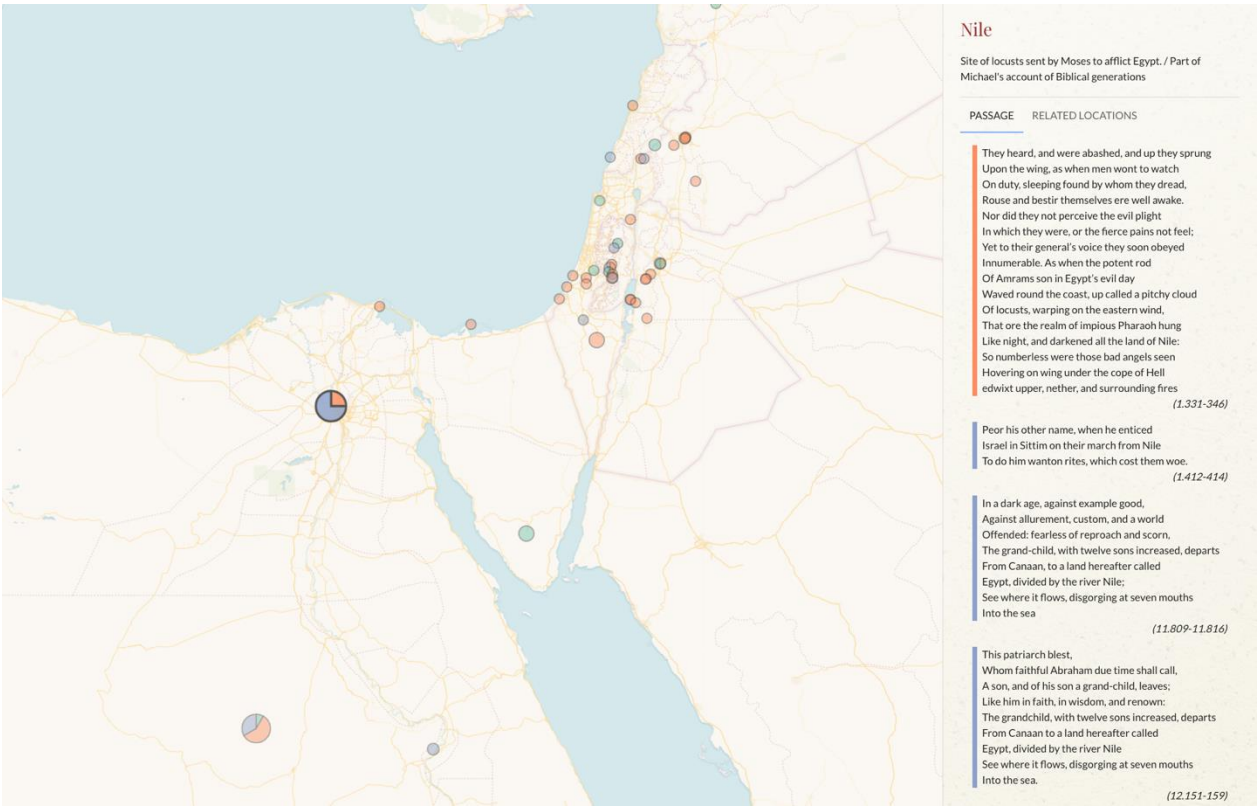


Figure 3

Pie charts show the context and passage for every separate place name reference if a place is mentioned more than once.

This prototyping moment is not meant to present a novel argument about *Paradise Lost*; as demonstrated through scholarly examples, snippets of the connection between platial references and moral valence exists in Milton studies, dispersed across many works. What we have done is approach each place in *Paradise Lost* through this close reading and research framework and visualize it on a map. In turn, this grants readers easier access to the spatiality of *Paradise Lost* and encourages them to seek patterns and make connections. Moral valence animates the places on the map in a dynamic interface, like Milton does through various critiques in the text, making it a useful resource for exploring and generating research questions. One can imagine visualizing other literary frameworks or arguments arrived at through close reading, for example, or including more georeferenced maps to study them against.

Prototyping moment 2: Parallel map visualizations

Navigating multiple works at once has never been easier—a stark comparison is the labour of using a bookwheel, once a groundbreaking sixteenth century-invention by Agostino Ramelli in the form of a rotating bookcase to facilitate an easier way of reading more than one large book at once, to tools like *Juxta* (McGann 2012), a collation tool used for comparing multiple editions on a single screen. The message across these technologies, however, is clear: there is scholarly value in comparative work. *A Map of Paradise Lost* has multiple layers; for example, we have manually geoparsed (identified all locatable place name references and matched them with their corresponding coordinates) the references to places in Genesis as a separate layer (purple), and included a layer of all places mentioned in the bible from *OpenBible* as an additional layer (white),

in an attempt to study the extent to which biblical naming was reproduced in *Paradise Lost* (Crossway 2011). However, maps have to be readable in order to make sense of them. Crowding all of these layers onto a single interface may be confusing, and overlaying historical maps simply counterproductive. In the case of *Paradise Lost*, more traditional methods of visualizing temporal change such as through timelines or annotations is not applicable—content that deals with biblical accounts, classical antiquity, mythology, and Milton's contemporary word, that oscillates between old geographical naming and new—does not lend itself to such defined temporal categories. Georectifying historical maps that, although imperfectly, but more closely capture these broad spatial categories is a more productive solution for this type of project. In an attempt to learn more about the geography that Milton references and the surrounding areas in these different contexts, and to more legibly navigate the multiple layers, we have built a functionality that allows to split the screen into up to four parts, where each part can be customized separately (Figure 4).

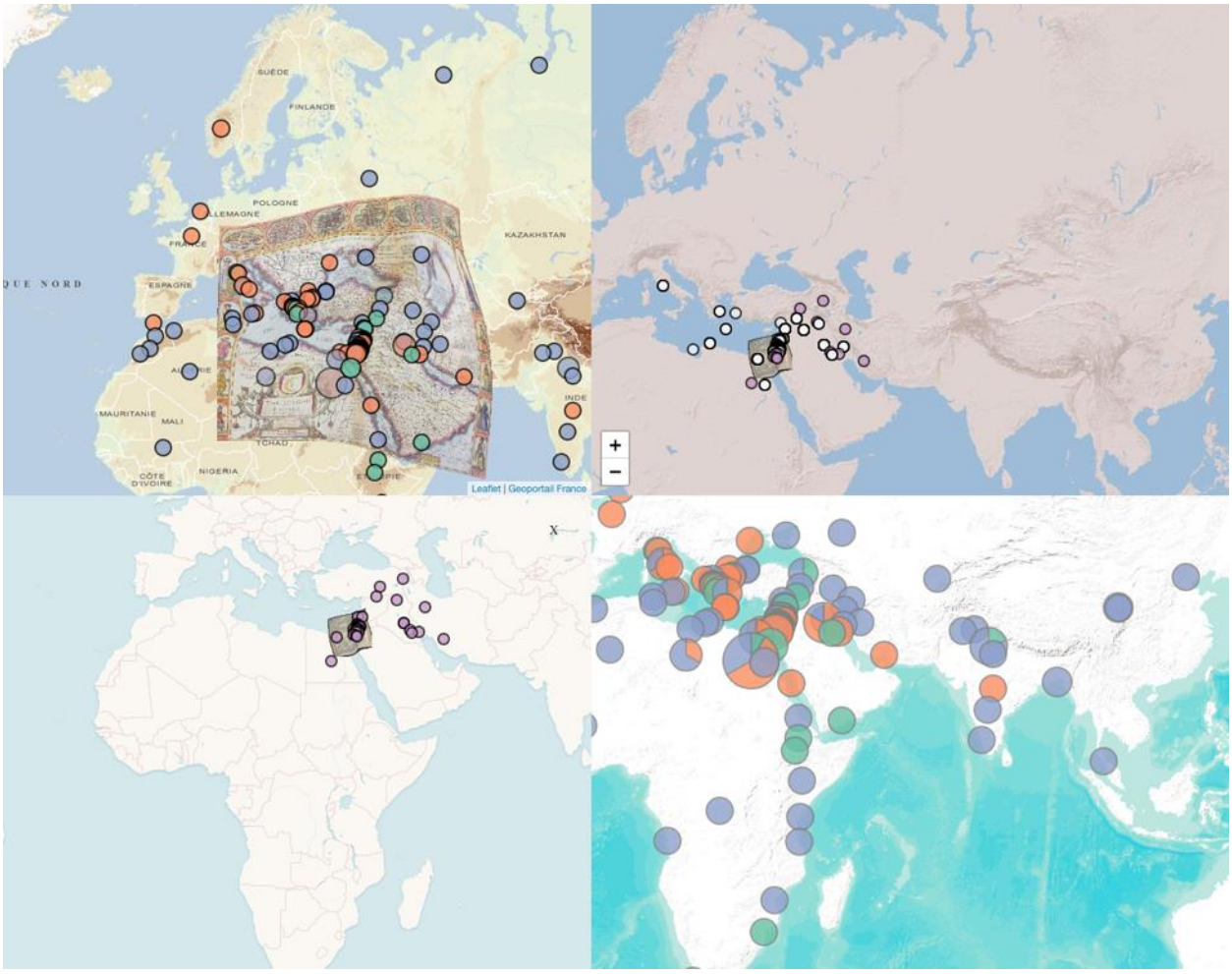


Figure 4

A Map of Paradise Lost parallel map visualizations.

Different basemaps, historical maps, layers, and marker types can be visualized and compared simultaneously, since the maps are synchronized. Some questions this feature can generate are: What are the same place names mentioned across different works? How globally encompassing are they, and where do they focus on most? How did place names and boundaries change over time? One can imagine that with additional

georectified layers, more historical comparison can be carried out. Even outside of the context of the epic poem, users can inquire how, for example, places were renamed after the Turkish conquest in comparison to earlier or current naming. By visualizing these maps side-to-side, parts of biblical and historical geography of the Levant, and place name referencing across the entirety of *Paradise Lost*, Genesis, and the Bible can be situated and understood by a broader audience.

Prototyping moment 3: Building capacity for interdisciplinary collaboration in mapping projects

The process of creating the data for geospatial humanities projects, including defining the categories and variables, extracting and cleaning data, standardizing it across, and reshuffling categories for generating the most readable and useful geovisualizations, is a meticulous process that requires revisiting the data many times, especially since we are carrying out all these steps manually for the sake of accuracy, since the content of *Paradise Lost* does not easily lend itself to automatic methodologies. In interdisciplinary collaborations, all parties bring their skillsets and the division of labour happens with respect to individual expertise. This means that updating, revising, and moving the data to fuel the map in a project like *A Map of Paradise Lost*, would necessarily require the humanist and developer to both be actively engaged in a stepwise process that may not always be a productive division of labour, and can actually impede the project from expanding in the future without the continuous active involvement of a developer in part of the process. Our final prototyping moment grants non-developers more autonomy in contributing to the project, while also planning for longer-term sustainability without having to rely on too many outside variables that increase the need for updates and iterative control. The solution to both aforementioned considerations is the data pipeline (see Figure 5).

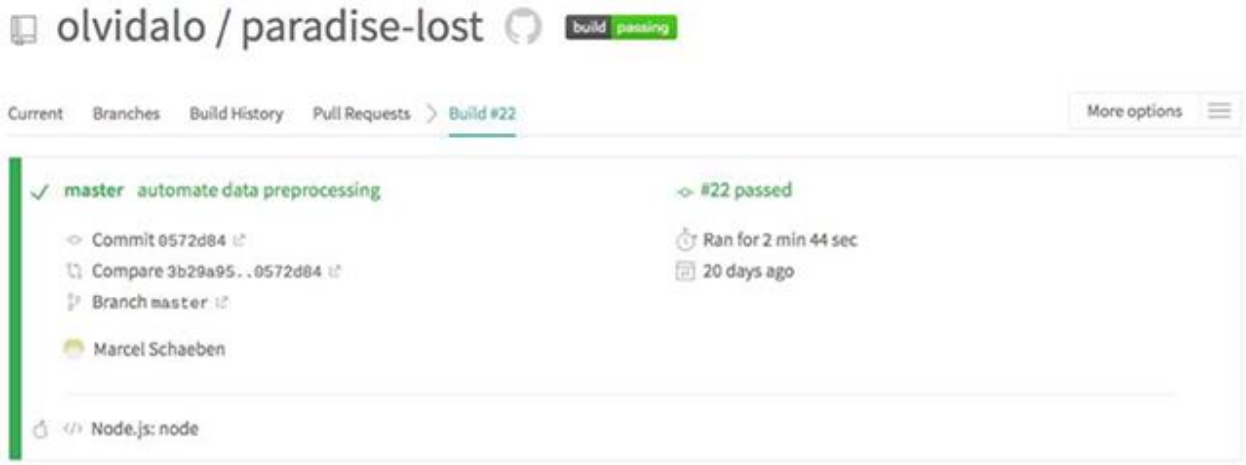


Figure 5

Travis CI running automated tasks on a remote server and pushing all the changes from the GitHub code and data repository to the GitHub Pages web hosting service for immediate online publication to *A Map of Paradise Lost*.

Organized in a single spreadsheet, the data platform is a building block of the application. All required processing of the data is carried out using the standard build tools that are used in the development of the app; we do not use an extra database, an extra server where the database lives, or entry forms. The idea of the pipeline is that any collaborator can edit the content with a knowledge of how to use spreadsheets and the very basics of the git-based collaborative platform GitHub. If a team member makes a change to a spreadsheet

and makes a commit to push the changes to GitHub, this sets off the pipeline which builds the app, validates if the data is complete and in the right format, and then publishes it to the web server. Through validation, ideally, nothing can be broken in the process. Humanists do not need to engage with the building process, pipelines, and other technical details that have already been developed, and can focus on the content. This division of labour raises the questions brought up by scholars such as Stephen Ramsay (2016) about whether a digital humanist has to be a coder or not; for this paper and project, this inquiry is rather narrow since the iterative nature of the project invites digital humanists, but also early modern scholars, to share their expertise on Milton by suggesting other potential close readings and literary frameworks for interpreting the spatiality of *Paradise Lost*. By making it a more straightforward task to contribute to the project, we are working towards a less labour-intensive, steep-learning-curve model that is more productive and accessible. Rather than insist that all contributors must have advanced coding skills and a background in early modern literature in order to equally contribute to all aspects of the project, we build on one of digital humanities' strongest suits: that through interdisciplinary collaboration, contributors from different fields can, together, build something neither could have done separately. The collaborative feature presented in this paper can open the project to a larger group of Milton scholars who can expand existing readings by adding new interpretive layers or editorial analyses from other scholarly editions.

Conclusion

By addressing the prototyping process of *A Map of Paradise Lost*, we sought to offer an explanation of select prototyping moments in order to address the question "why map literature?"—essentially pointing to how the process of mapping is not unlike that of close reading, and that through data gathering and visualization, existing and novel interpretations of literature can be validated, expanded, and contested. The digital medium also encourages a space for creativity and for experimentation with scholarly communication and methodologies; for example, the project introduces a novel approach for cartographic comparison where maps and data layers can be visualized side-by-side for more intuitive explorations. The data pipeline provides a model for collaboration in which contributors with different technical skills can more readily contribute to a project, in an attempt to encourage a community of practice among literary experts that can essentially be adapted to other projects and their respective content. In doing so, we are also conscious of the iterative nature of digital projects and their ephemerality in terms of maintenance, thus thinking and prototyping towards sustainability.

Acknowledgements

A note of gratitude to Dr. David Currell, who worked with Randa El Khatib on a separate chapter on *A Map of Paradise Lost* that engages the project in an early modern context, and to Dr. Øyvind Eide for facilitating this collaboration.

WORKS CITED:

Campbell, Gordon. 2007. "To the Shore of Tripoli': Milton, Islam, and the Attacks on America and Spain." In *Fundamentalism and Literature*, edited by Catherine Pesso-Miquel and Klaus Stierstorfer, 7-19. New York: Palgrave Macmillan.

Caquard, Sebastién. 2011. “Cartographies of Fictional Worlds.” *The Cartographic Journal* 48 (4): 224–5. doi: 10.1179/174327711X13190991350051.

Cooper, David, Christopher Donaldson, and Patricia Murrieta-Flores, eds. 2016. *Literary Mapping in the Digital Age*. New York: Routledge.

Cooper, David and Ian Gregory. 2011. “Mapping the English Lake District: A Literary GIS.” *Transactions of the Institute of British Geographers* 36 (1): 89-108. doi:10.1111/j.1475-5661.2010.00405.x.

Cresswell, Tim. 2015. *Place: An Introduction.* Hoboken, NJ: Wiley-Blackwell.

Crossway Bibles. 2011. OpenBible.info. *Good News Publishers*. Accessed July 15, 2019. https://www.openbible.info/.

El Khatib, Randa, and David Currell. 2018. “Mapping the Moralized Geography of *Paradise Lost*.” In *Digital Milton*, edited by David Currell and Islam Issa, 129-52. Cham,Switzerland: Palgrave Macmillan.

Evans, Martin, Kenneth Ligda, Karl Grossner, and David McClure. 2011. *Authorial London*. Stanford University. Accessed July 15, 2019. https://cidr-authorsial-prod.stanford.edu/.

Fuller, Thomas. 1662. Pisgah-Sight of Palestine and the Confines Thereof: With the History of the Old and New Testament Acted Thereon. London: Printed by J. F. for John Williams.

Galey, Alan, and Stan Ruecker. 2010. “How a Prototype Argues.” *Literary and Linguistic Computing* 25 (4): 405–24.

Heylyn, Peter. 1657. *Cosmographie in Four Books: Containing the Chorographie and History of The Whole World, and All the Principall Kingdomes, Provinces, Seas, and Isles Thereof*. London: Anne Seile and Philip Chetwind.

Liu, Alan. 2013. "The Meaning of the Digital Humanities." *PMLA* 128 (2): 409- 23.

Luchetta, Sara. 2018. “Going Beyond the Grid: Literary Mapping as Creative Reading.” *Journal of Geography in Higher Education* 42 (3): 384-411. doi: 10.1080/03098265.2018.1455172.

Milton, John. 2007. *The Complete Poetry and Essential Prose of John Milton*, edited by William Kerrigan, John Rumrich, and Stephen M. Fallon. New York: The Modern Library.

Lim, Walter S. H. 2010. “John Milton, Orientalism, and the Empires of the East in *Paradise Lost*.” In *The English Renaissance, Orientalism, and the Idea of Asia*, edited by Debra Johanyak and Walter S. H. Lim: 203-35. New York: Palgrave Macmillan.

Map of Biblical Lands. 1611. King James Bible. Rare Book and Manuscript Library, University of Pennsylvania. Woodcut. 41 cm (folio).

McGann, Jerome. 2012. Juxta. *Applied Research in Patacriticism*. Accessed on July 30, 2019. http://www.juxtasoftware.org/.

McColley, Grant. 1937. "The Epic Catalogue of *Paradise Lost*." *ELH* 4 (3): 180-91. doi:10.2307/2871532.

Muri, Allison. 2016. “Beyond GIS: On Mapping Early Modern Narratives and the Chronotope.” *Digital Studies/ Le champ numérique* 6. doi: 10.16995/dscn.11.

Muri, Allison, et al. 2016. *The Grub Street Project*. Digital Research Center: University of Saskatchewan.

Ng, Morgan. 2013. “Milton’s Maps.” *Word and Image* 29 (4): 428-42.

Piatti, Barbara. 2016. “Mapping Fiction: The Theories, Tools and Potentials of Literary Cartography.” In *Literary Mapping in the Digital Age*, edited by David Cooper, Donaldson, Christopher and Patricia Murrieta-Flores, 88–101. New York: Routledge.

Piatti, Barbara, and Lorenz Hurni. 2011.” Cartographies of Fictional Worlds.” *The Cartographic Journal* 48 (4): 218-23. doi: 10.1179/174327711X13190991350051.

Prickman, Greg, Andrew Holland, Robert Shepard and Steve Tomblin. 2013. *The Atlas of Early Printing.* The University of Iowa Libraries.

Quint, David. 2014. *Inside Paradise Lost: Reading the Designs of Milton’s Epic*. Princeton: Princeton University Press.

Ramsay, Stephen. 2016. “Who’s In and Who’s Out.” In *Defining Digital Humanities: A Reader*, edited by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte, 239-241. New York: Routledge.

Ruecker, Stan. 2015. “A Brief Taxonomy of Prototypes for the Digital Humanities.” *Scholarly and Research Communication* 6 (2): 1-11.

Sandys, George. 1610. *A Relation of a Journey Begun An: Dom: 1610. Fovre bookes. Containing a Description of the Turkish Empire, of Ægypt, of the Holy Land, of the Remote Parts of Italy, and Ilands Adioyning*. London: Printed for W. Barrett.

Sahle, Partick. 2016. “What is a Scholarly Digital Edition.” *In Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 19-40, Cambridge, UK: Open Book Publishers. doi: 10.11647/OBP.0095.

Speed, John. 1626. “The Turkish Empire. In A Prospect of the Most Famous Parts of the World.” Maps & Imagery Library, Special and Area Studies Collections, George A. Smathers Libraries, University of Florida.

Viktus, Daniel. 2003. *Turning Turk: English Theater and the Multicultural Mediterranean, 1570-1630*. New York: Palgrave Macmillan.