Journal of Open Humanities Data

The Curated *UNESCO Courier* 1.0: Annotated Corpora for Digital Research in the Global Humanities

RESEARCH PAPER

BENJAMIN G. MARTIN
FREDRIK MOHAMMEDI NORÉN
ROGER MÄHLER
ANDREAS MARKLUND
ORIANE MARTIN

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The monthly magazine of the United Nations Educational, Scientific and Cultural Organization, founded in 1948 as *The UNESCO Courier*, represents an extraordinary resource for research on global themes in the humanities. We present the Curated Courier 1.0, a package of digital text corpora, text analysis tools, and supplementary material that aims to make the complete archive of this publication from 1948 to 2020 machine-readable, accessible, and reusable for digital text analysis. One corpus compiles the text of all articles, which we carefully reconstructed and linked to a comprehensive curated metadata index while excluding additional text (masthead, photo captions, letters to the editor, and so on). A second corpus brings together the complete text of all issues. This article first presents the value of *Courier* as a source for digital research in the global humanities. Second, it outlines how we created the curated corpus and discusses some challenges we met. Third, it offers examples of tools researchers might use to explore and utilize the annotated corpus and discusses a few approaches that we have developed and tested.

CORRESPONDING AUTHOR:
Benjamin G. Martin
Department of History of Science and Ideas, Uppsala University, Uppsala, Sweden
benjamin.martin@idehist.uu.se

KEYWORDS:
UNESCO; international organizations; history; global humanities; text analysis; topic modeling

TO CITE THIS ARTICLE:
Martin, B. G., Norén, F. M., Mähler, R., Marklund, A., & Martin, O. (2024). The Curated *UNESCO Courier* 1.0: Annotated Corpora for Digital Research in the Global Humanities. *Journal of Open Humanities Data*, 10: 20, pp. 1–13. DOI: https://doi.org/10.5334/johd.181

(1) CONTEXT AND MOTIVATION

(1.1) INTRODUCTION

Many disciplines in the humanities have experienced a "global turn," marked by a growing focus on "contemporary and historical processes of globalization" and by increased engagement "with scholars beyond the Euro-American academy" (Darian-Smith & McCarty 2017: 2; see also Amirell 2023). This trend has been echoed by heightened curiosity about the mechanisms of global intellectual and cultural life, for example, through increased interest among humanities scholars in international organizations, including both nongovernmental organizations and intergovernmental organizations like the United Nations (early examples include Iriye 2002; Kott 2011). For some scholars, studying these global forums is part of a broader effort to pay more attention to voices from beyond the West and explore how people outside the West have shaped the modern world order. Many of the questions raised by scholars working in these ways call for global textual sources—sources that give us access to voices of individuals from around the world and/or that we know were read by a broad, international audience.

One body of text that meets these challenges is the monthly magazine of the United Nations Educational, Scientific and Cultural Organization (UNESCO), *The UNESCO Courier*. Founded in 1948, this magazine, usually referred to simply as *Courier*, featured articles by prominent thinkers from around the globe and appeared in its heyday in over thirty languages. During the Cold War, it was one of few publications distributed on both sides of the "Iron Curtain" and throughout the global south. Throughout, *Courier* addressed many themes of interest to humanities scholars in such fields as literature, music, history, educational theory, archaeology, philosophy, anthropology, and the arts.

Courier thus represents a rich resource and a rather extensive dataset, comprising (for the period 1948–2020) some 13 million words. It is, put differently, just the sort of source that would be of great utility in digital form. In recent years, UNESCO has indeed made the magazine's earlier issues accessible online in the form of PDF files.[1] These files make it possible for users anywhere to read individual issues. Still, they do not allow for full-text searching, much less any of the computational text analysis methods that have recently contributed to essential advances in humanities research. As part of the research project "International Ideas at UNESCO," our team has sought to address this issue: we have assembled and curated all issues of the English-language edition of the *UNESCO Courier* from 1948 to 2020, generating a package of machine-readable text corpora we call the *Curated Courier 1.0*. This includes: (1) the annotated article corpus; (2) the complete issue corpus; (3) a suite of analytical tools in Jupyter Notebook format; and (4) supplementary materials, such as documentation and quality control files. Our goal in creating this package has been to make this valuable global source accessible for new forms of digital research.

In this article, we present the Curated *Courier* 1.0 and discuss its potential uses in three sections. The first section offers a brief historical discussion of *Courier*, focusing on its value as a source for digital research in the global humanities. The second section presents the process of creating the curated corpus, outlines its technical features, and discusses some challenges we met in producing it. In the third section, we offer several examples of tools researchers might use to explore and utilize the annotated corpus. We also discuss a few approaches that we have developed and tested.

(1.2) *COURIER*: A RESOURCE FOR RESEARCH IN THE GLOBAL HUMANITIES

If there has ever been such a thing as a global magazine, it may have been *The UNESCO Courier*. Founded in 1948, its mission was to "promote UNESCO's ideals, maintain a platform for the dialogue between cultures, and provide a forum for international debate."[2] Initially formatted like a newspaper, *Courier* served UNESCO first as a tool for informing readers worldwide about the organization and its activities. By the mid-1950s, in conjunction with its transition into a monthly illustrated magazine format, it became an "opinion journal tackling topics covered by UNESCO's mandate"—that is, education, science, culture and communications—"but with little reference to activities carried out or official points of view" (Defourny 2003: 428–429).

1 https://courier.unesco.org/en/archives (accessed 19 January 2024).

2 https://courier.unesco.org/en/about (accessed 19 January 2024).

DOI: 10.5334/johd.181

The editorial staff—led until 1977 by the American journalist Sandy Koffler and in the 1980s by the prominent Martiniquais poet and writer Édouard Glissant—wanted the publication to serve as a forum for debate and self-consciously strove to give voice to representatives of countries around the world. Among the scholars, scientists, philosophers, literary writers, and other intellectuals who published in *Courier*'s pages, we find figures ranging from the South African novelist Nadine Gordimer to the Soviet composer Dmitri Shostakovich, from the French anthropologist Claude Lévi-Strauss to the Indian filmmaker Satyajit Ray, from Swedish social thinker and diplomat Alva Myrdal to Senegalese poet and president Léopold Sédar Senghor, and from Canadian media theorist Marshall McLuhan to Swiss physicist Albert Einstein. Published first in English and French, *Courier* soon added an edition in Spanish, and in the early 1960s added editions in Arabic, German, Italian, and Japanese. Between 1967 and 1973, *Courier* also began to appear in Dutch, Hebrew, Hindi, Persian, Portuguese, Tamil, and Turkish. The magazine reached a peak of linguistic coverage in 1988 when it appeared in thirty-five languages (Dausse 2018: 70).

Given the range of voices it has included and the international spread of its readership, *Courier* is a resource of particular interest for research in what has been called the "global humanities" (Amirell 2023). This refers to humanities research that seeks "to take on the study of cultural and historical processes without geographic or chronological limitations and unfettered from the bonds of Eurocentrism and methodological nationalism, which often are associated with the traditional humanities disciplines as they have developed in the West" (Amirell 2023: 2). There are, of course, limitations to the magazine's value as a source for such research. *Courier* has by no means been free from Eurocentrism; this is evident from some of the attitudes expressed in its pages and the disproportionate frequency of Western voices even after the 1960s. UNESCO is, moreover, an organization of states, and its magazine has sometimes had a corresponding tendency to emphasize the principle of nationality, even in its efforts to promote cross-cultural understanding. Nonetheless, *Courier*'s value as a global source has been demonstrated by the scholarship that has already made use of it. Such research includes work by scholars in literature study, history, heritage studies, archaeology, musicology, and research on UNESCO and its programs, which has expanded greatly in recent decades.[3] Moreover, *Courier* is, as a publication, an important research object in its own right. The scholarly literature about *Courier* itself is limited (Defourny 2003; Simonsen 2020; Krebs 2016). But insofar as it offers an example of a self-consciously global magazine, *Courier*'s strategies, language, and style offer material for studying global media history.[4]

The ability of researchers to make the most of the resource that *Courier* represents has been limited by the publication's vast scale. Our goal in creating the Curated *Courier* 1.0 has been to make it possible to approach this rich material using methods of textual analysis associated with the digital humanities.[5] The reason to do this is not simply because the hundreds of *Courier* issues published between 1948 and 2020 are too much for one person to read; the point is also that digital tools allow scholars to approach the material in other ways than those that reading allows. Existing digital humanities research on international materials (e.g. Moretti & Pestre 2015) suggests that digital approaches may be particularly well suited to the kind of global source that *Courier* offers.

Moreover, *Courier* presents several advantages that recommend it for preparation as a set of digital corpora. To begin with, it includes a large number of words: the complete *Courier* (1948–2020) consists of some 13 million tokens. Second, compared to other large collections of texts, the collected issues of *Courier* form a thematically and institutionally coherent whole. The publication was broad in the range of topics it addressed but specific in its link to UNESCO and its mandate. Third, the journal has presented a high degree of continuity in its format over most of its history. From 1948 to 2001, it appeared in ten to twelve issues per year. Its layout has been consistent since the 1950s, and until 2001, the total amount of text data it included

3 Examples include work by literary scholars (Brouillette, 2019; McDonald, 2017) and cultural historians (Betts, 2020; Wong, 2008). The heritage studies community has shown much interest in UNESCO (Gfeller & Eisenberg, 2016; Meskell, 2018) as has scholarship on the history and politics of archaeology (Carruthers, 2022).

4 Regarding UNESCO's own role in advancing international media research, see Hamelink (2012).

5 The literature on these methods is very large; as an introduction see Moretti (2013); Underwood (2019); Guldi (2023).

per year remained relatively stable, although increasing from the 1980s (Figure 1).[6] Finally, unlike commercial publications, *Courier* is not copyright protected; UNESCO has made the text available for public use, including (for the period since 2017) through a Creative Commons license.[7]

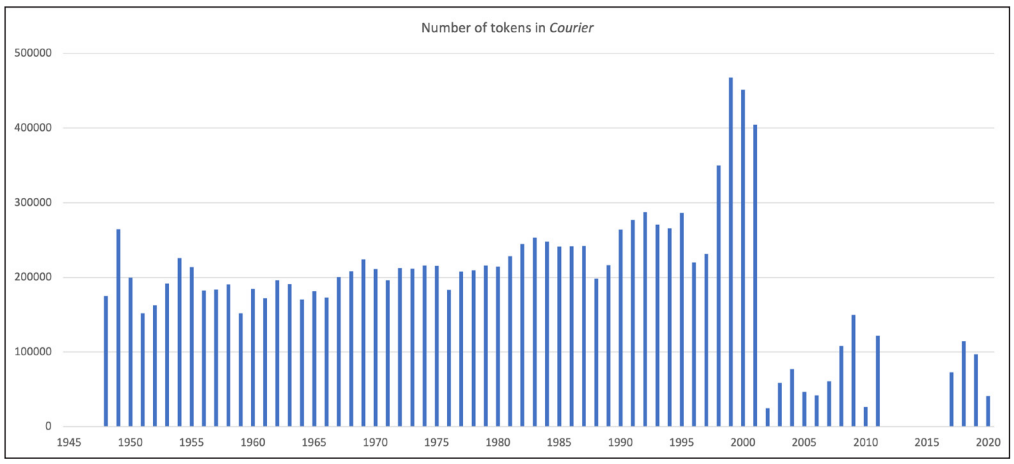


Figure 1 The number of tokens per year in *Courier* (non-curated). A token is a compound of characters separated by spaces.

(2) DATASET DESCRIPTION

OBJECT NAME

The Curated Courier 1.0

FORMAT NAMES AND VERSIONS

Text corpora: .txt

Metadata: CSV

CREATION DATES

2021-02-01 to 2023-06-22.

DATASET CREATORS

Andreas Marklund (developer, Humlab, Umeå University), Benjamin Martin (PI, Uppsala University), Oriane Mathilde Martin (assistant data curator, Uppsala University), Fredrik Mohammadi Norén (researcher, Malmö University), and Roger Mähler (developer, Humlab, Umeå University).

LANGUAGE

English

LICENSE

Creative Commons Attribution 4.0 International (CC BY 4.0 Deed).

REPOSITORY NAME

Zenodo (DOI: 10.5281/zenodo.10083489) [https://doi.org/10.5281/zenodo.10083489]

PUBLICATION DATE

2023-11-08

6 From 2002 the rate of issues per year became irregular: publication ceased entirely between 2012, when only one issue appeared, and April 2017, when the magazine was relaunched at a rate of four issues per year, appearing in UNESCO's six official languages (Arabic, Chinese, English, French, Russian, and Spanish), as well as occasionally in Catalan and Esperanto. https://courier.unesco.org/en/about (accessed 19 January 2024).

7 See: https://courier.unesco.org/en/about (accessed 19 January 2024).

(3) METHOD

(3.1) CREATING THE CURATED COURIER 1.0

The starting point for preparing a machine-readable corpus of *Courier* was the PDF files of each issue made available by UNESCO in the English-language edition.[8] We chose to prepare digital corpora of the English-language version, believing this to have the most comprehensive reuse value. (The methods we outline here can be applied in the future to the preparation of parallel corpora derived from *Courier*'s other editions.) We scraped these PDFs into text files, one per issue. However, the heart of *Courier* is the articles published in the magazine. Hence, one of the main ambitions of the Curated *Courier* 1.0 package, designed to enhance its reusability, was to provide researchers with a corpus consisting solely of the articles published in *Courier* from 1948 to 2020. Isolating and compiling these texts proved a substantial challenge, however. Like most historical text collections, *Courier* presents "non-standard data," which requires careful preparation before it is usable for research through digital methods (Mäkelä et al. 2020). This section describes the curating process—inspired by the ideal of agile corpus creation (e.g. Voormann & Gut 2008) and collaborative curation (Mueller & Burns 2016). As illustrated in Figure 2, the curating work included two main tasks: (1) scraping all *Courier* issues in PDF formats from UNESCO's web archive to extract the OCRed text from the PDFs, and (2) identifying, assembling, and annotating the texts of all articles. This process included marking up the text files to link them to metadata for article title, author(s), year, and topic tags, as well as assessing quality assurance dimensions of the corpus.

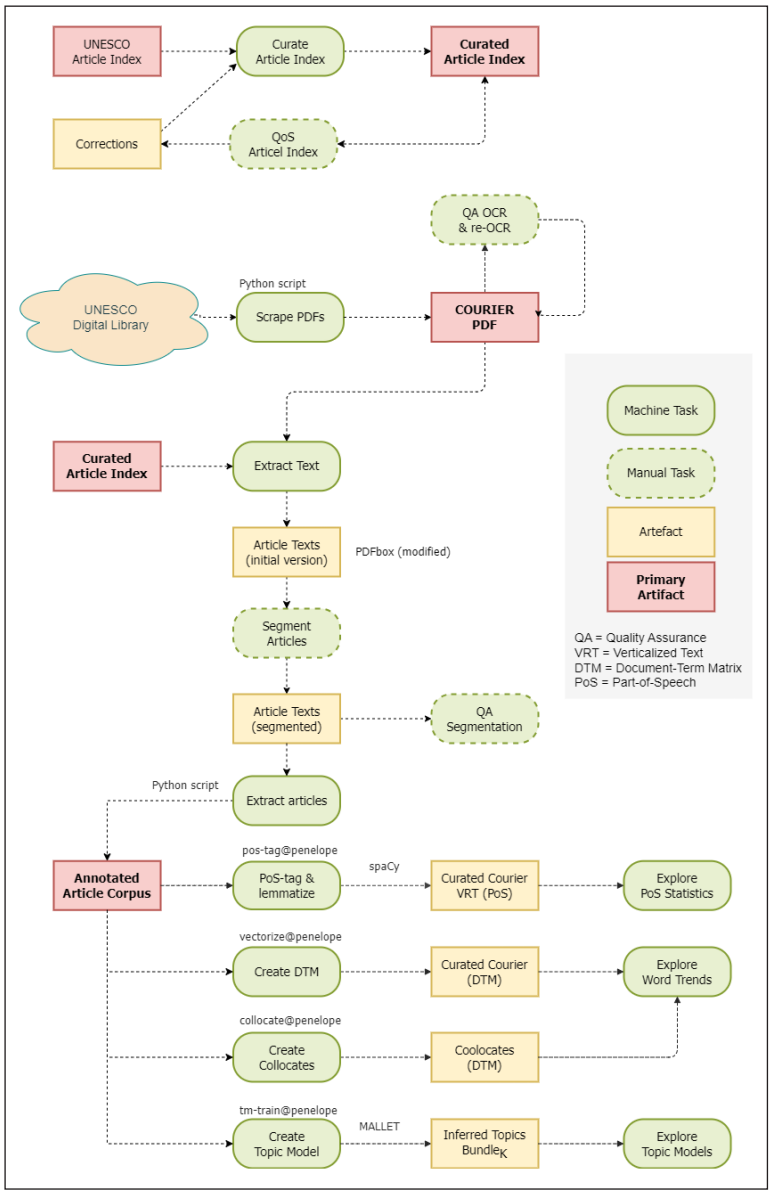


Figure 2 Courier curation workflow: This maps the processes applied to create the Curated Courier 1.0 package.

8 These PDFs can be accessed via the *Courier* online archive (https://courier.unesco.org/en/archives), but also directly through UNESCO's online library, UNESDOC: https://unesdoc.unesco.org (accessed 19 January 2024).

(3.1.1) Scraping and extracting *Courier* text data

We used a custom Python script to scrape all the *Courier* issue PDF files. Extracting OCRed text from a PDF file is a complex task. For example, while OCR systems have greatly improved over the years, they can still encounter difficulties in accurately recognizing text and capturing the correct ordering of the text (Jarlbrink & Snickars, 2017). The latter issue becomes particularly significant when dealing with complex layouts, such as the *Courier*'s changing graphical design. Since the OCR process results in a set of text blocks positioned on each page, the challenge boils down to sorting these blocks into a correct read order (e.g. Hurtado Bodell, Magnusson & Mützel 2022). This sorting can be especially problematic when a page has multiple columns (earlier *Courier* issues have up to five columns). Furthermore, text segments on a page might belong to more than one article, constituting margin notes and image captions. Text can also be part of a centerfold where the article spans across two pages. Hence, extracting text from PDF risks merging different text columns or misplacing their reading order.

In the process of choosing a text extraction tool, one also has to decide whether the page text should be presented as a mirrored version of the digitized page image or if the text should reflect how a human usually reads: first the page header, then body text from top to bottom and column after column, ending with page footer. The latter was chosen for the curated *Courier* corpus. Moreover, each PDF format stores text in its own way. For example, PDF is biased towards page layout; since a single word may contain several textual elements due to different design choices (e.g. the size, font, and color), text might be encoded using different standards (e.g. Déjean & Meunier 2006). The following tools were tried out and compared for extracting OCRed text from *Courier*: MS Word, Pdfbox, Pdfminer, Pdftotext, Pdfplumber, and Tesseract.[9] In this study, patterns of poor tool quality were identified in a random sample of pages drawn from different decades. The results indicated that Pdfbox and Tesseract outperformed MS Word, Pdfminer, Pdfplumber, and Pdftotext in terms of effectively extracting text in the correct read order. Pdfbox (Apache PDFBox, 2023) was chosen for text extraction, and Tesseract (e.g. Smith, 2007) for re-OCR of some twenty *Courier* pages that we found had unsatisfactory OCR quality.

(4) RESULTS AND DISCUSSION

Annotation and segmentation of digitized historical newspapers are notorious for the challenges and obstacles they present (Ehrmann et al. 2020; Barman et al. 2021). In our case, this work first required isolating each article and excluding extraneous textual material. Next, we needed to consolidate the text of each article; this was complicated by the fact that a single article sometimes stretched over several contiguous pages but then continued and ended much later in the same issue. The changing layout structure of *Courier* articles also posed various difficulties. For example, the magazine has not used article headlines or marked the end of an article in a consistent way. For these reasons, finding rules to automatically identify the start and end of articles was not trivial. This task required special attention and was carried out both using automatized and manual approaches in an iterative fashion (cf. Voormann & Gut 2008).

(4.1) ANNOTATING *COURIER* ARTICLES

The point of departure for our annotation work was based on an article index produced by UNESCO (and generously shared with us): this index listed 7612 English-language articles for the period 1948–2020, providing detailed metadata about the article title, author name, translated languages, issue number, volume, and year. Another metadata feature is the page number on which each article starts and ends, including page numbers for articles that occur on multiple but non-contiguous pages. We used this page data from the article index to identify the start and end of individual articles through a two-step, semi-automatic approach (cf. Ren & Matsumoto 2016). Using the page metadata, and experimenting with identifying the beginning of articles by searching for different font sizes, we automatically filtered out pages that lack any indexed article text altogether (which accounted for 17 % of all pages).[10]

9 Our selection process is detailed at: https://github.com/inidun/unesco_data_collection/issues/5 (accessed 19 January 2024).

10 Details on this process are available at: https://github.com/inidun/tagged_courier/blob/main/workflow/README.md (accessed 19 January 2024).

Based on the article index, we determined that over 600 pages contained both the end of one article and the beginning of another. We designed and implemented a tailor-made process of manual curation to improve the article segmentation for such conflicting pages. Each *Courier* page, which potentially contained two articles, was reviewed carefully. To do this, we compared the text file produced by our automatized article segmentation to the relevant images in the original *Courier* PDFs. Then, we annotated the text file by adding several tags or labels to the text to facilitate automated compiling, for example, by inserting the correct article IDs in the correct positions. Sometimes, textual elements were also identified that did not belong to the article, such as advertisement-like content between two articles, for example, information about newly published books. Other challenges regarded determining whether certain text segments were part of the article or not. Sometimes, for instance, a shorter text with a different layout discussed the same subject by the same author as the indexed article close to this shorter text segment. Hence, knowledge of the broader layout changes throughout the history of the magazine was crucial during the manual annotation work. In this part of the curation process, some scenarios required special attention. For instance, specific labels were used if the text belonged to an editorial statement or if an article was missing in the code. Similarly, issue supplements, denoting content not typically part of the regular *Courier* issue, were given their own tag.

(4.2) QUALITY ASSURANCE

Assessing the quality of curating work is pivotal (Hurtado Bodell et al., 2022). We evaluated the quality of the annotated *Courier* article corpus in two dimensions: the OCR quality, and the quality of our process of annotating and compiling the *Courier* articles. To assess OCR quality, we selected two random text chunks of approximately 100 words from two different articles per year (1948–2020). Each chunk was extracted from the PDF and pasted into a word processing program to manually calculate the number of OCR errors per unit of text. On average, this test showed a 0.7 % error rate, which should be considered good quality. The article annotation quality control then examined different quality dimensions: identifying whether or not the extracted article texts begin and end in the correct place (confirming, that is, that we had the complete texts of all articles), identifying any excess textual data in annotated articles, and determining that the text columns appeared in the correct order. Here, we randomly selected a sample of two articles per year. Based on this sample, we found that articles were incomplete in 8% of cases. However, the errors thus identified were minor: they were mostly cases of a missing caption (5 of 11 error cases) or poor OCR (3 of 11), and only rarely missing paragraphs (3 of 11). Excess textual data—text that does not belong to the article—was present in 6% of the sampled articles. Most of the time, however, this excess consisted of minor segments like figure captions or introductory text belonging to other articles. These results can be considered as quite good. Ensuring the correct ordering of columns was somewhat more difficult: 15% of the sampled articles were found to include an error in column order. Nonetheless, for many forms of digital analysis these errors in column order represent only a small problem: methods such as word trends and topic modeling are, after all, dependent on word distribution rather than word order.

Finally, we used the article annotations, produced through both our manual and automated processes, to compile a curated corpus consisting (only) of articles published in *Courier*. A custom-made Python script was used to extract and merge all article segments into a single corpus document. The script also checked for potential errors in the extracted articles, such as mismatching years, duplicated record numbers, missing articles, and articles with no pages. In the process of curating the articles we also produced an enhanced article index, or metadata file, which includes multiple such additions and corrections.

(5) IMPLICATIONS/APPLICATIONS

(5.1) REUSE POTENTIAL

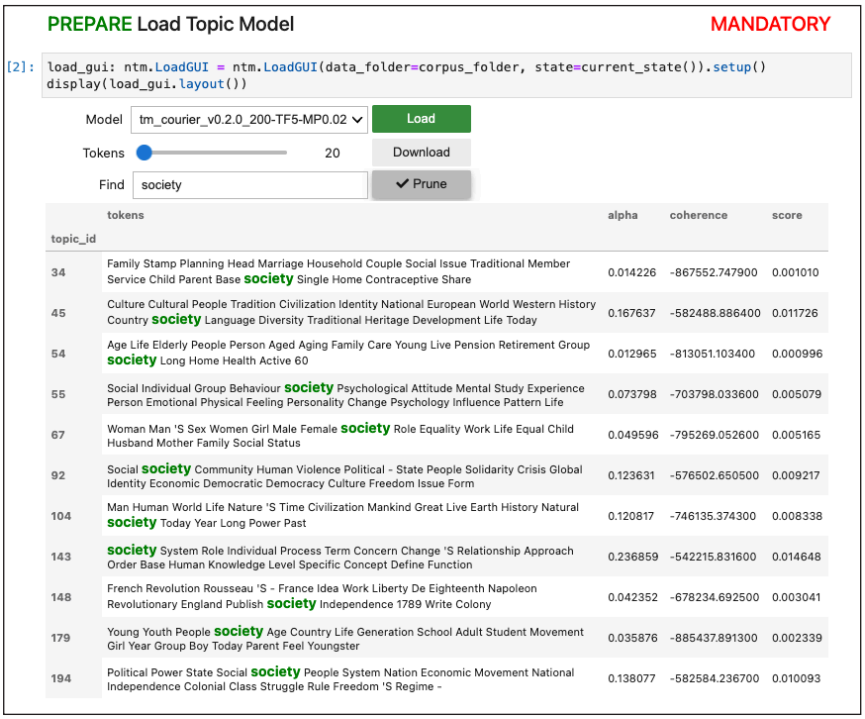
The Curated *Courier* 1.0 package contains multiple features with reuse potential. The main items are the two text corpora: the annotated article corpus (1948–2020) and the complete issue corpus (1948–2020), both of which are available on Zenodo (https://zenodo.org/records/10083490) and on GitHub (https://github.com/inidun/curated_courier). At our project site on GitHub, we have also stored supplementary materials: these include the complete issue corpus downloadable as .txt files (one per issue) along with links to stored PDF versions of each *Courier* issue, as well as documentation and quality control files.

In addition, we have made available a suite of computational text analysis tools in a Jupyter notebook format. We developed these tools in order to conduct research on the *Courier* material in parallel to (and in support of) our work in creating the annotated article corpus. This parallel process made it easier to detect flaws in the text data and to identify what was needed in the curating work from a research perspective. By making these tools available now, we aim to enhance the reuse potential of the annotated article corpus. The tools allow researchers to measure part-of-speech statistics, explore word trends, and analyze the corpus through topic modeling. For text analysis purposes, the curated article corpus was part-of-speech (PoS) tagged, using the spaCy PoS tagger.[11] Using PoS-tagging makes it possible to study how the relative usage of nouns, verbs, and adjectives has changed over time. Using the word trend tool in our Jupyter notebooks, researchers can measure the frequency with which a particular word—or the lemmatized form of the noun—has occurred in the corpus over time (in absolute or normalized numbers).[12]

(5.2) COMPUTATIONAL TEXT ANALYSIS TOOLS

One text analysis tool we have found particularly useful for exploring the annotated article corpus has been Latent Dirichlet Allocation (LDA) topic modeling. In what follows, we describe some features of the topic modeling tools that we used and suggest their analytical potential. Topic modeling is a probabilistic method suitable for structuring a large and diverse text collection to a pre-set number of themes called "topics." The model assigns a probability value to each word in each document (in this case, each curated Courier article) and orders them into blends of "topics". This process results in a top list of the most likely occurring words in each topic (Blei et al., 2012). We have used the popular LDA topic modeling method, as implemented in Mallet (Blei et al., 2003; McCallum, 2002).

In our GitHub repository, we have uploaded models of 50, 100, 200, and 500 topics derived from the annotated article corpus. These models make it possible to explore different analytical resolutions: fewer topics yield broader categories while more topics generate a fine-grain structure. If, for example, we are interested in studying notions of "society" in *Courier*, the four models offer different analytical perspectives from which to chart this concept in the magazine. Searching for the term "society" among the top 20 words of the algorithmically generated "topics" in the different models, one finds three such topics in the 50-topic model—encompassing rather broad themes having to do with state governments, world culture, and women and family. In the 200-topic model we find 11 topics in which the word "society" appears among the top 20 words, including themes similar to those found in the 50 model but also several others (Figure 3).



topic_id	tokens	alpha	coherence	score
34	Family Stamp Planning Head Marriage Household Couple Social Issue Traditional Member Service Child Parent Base **society** Single Home Contraceptive Share	0.014226	-867552.747900	0.001010
45	Culture Cultural People Tradition Civilization Identity National European World Western History Country **society** Language Diversity Traditional Heritage Development Life Today	0.167637	-582488.886400	0.011726
54	Age Life Elderly People Person Aged Aging Family Care Young Live Pension Retirement Group **society** Long Home Health Active 60	0.012965	-813051.103400	0.000996
55	Social Individual Group Behaviour **society** Psychological Attitude Mental Study Experience Person Emotional Physical Feeling Personality Change Psychology Influence Pattern Life	0.073798	-703798.033600	0.005079
67	Woman Man 'S Sex Women Girl Male Female **society** Role Equality Work Life Equal Child Husband Mother Family Social Status	0.049596	-795269.052600	0.005165
92	Social **society** Community Human Violence Political - State People Solidarity Crisis Global Identity Economic Democratic Democracy Culture Freedom Issue Form	0.123631	-576502.650500	0.009217
104	Man Human World Life Nature 'S Time Civilization Mankind Great Live Earth History Natural **society** Today Year Long Power Past	0.120817	-746135.374300	0.008338
143	**society** System Role Individual Process Term Concern Change 'S Relationship Approach Order Base Human Knowledge Level Specific Concept Define Function	0.236859	-542215.831600	0.014648
148	French Revolution Rousseau 'S - France Idea Work Liberty De Eighteenth Napoleon Revolutionary England Publish **society** Independence 1789 Write Colony	0.042352	-678234.692500	0.003041
179	Young Youth People **society** Age Country Life Generation School Adult Student Movement Girl Year Group Boy Today Parent Feel Youngster	0.035876	-885437.891300	0.002339
194	Political Power State Social **society** People System Nation Economic Movement National Independence Colonial Class Struggle Rule Freedom 'S Regime -	0.138077	-582584.236700	0.010093

Figure 3 Print screen image of the Notebook feature "Load Topic Model", showing the 11 topics that contain the word "society" among their 20 most likely words from a model of 200 topics.

11 https://zenodo.org/record/8123552 (accessed 19 January 2024).

12 The notebooks are available at: https://inidun.github.io/courier-lab (accessed 19 January 2024).

Using the Notebook feature "Topic-Word Distribution" it is then possible to study the probabilistic rank order of the words included in each topic. This tool assists the researcher in interpreting each topic and assigning them more precise labels. The graphs the tool produces show that it is often the top five words (or even fewer) that determine the characteristics of a topic (Figure 4). Here, as an example, we focus on topic 143 (from the 200-topic model): given that its most likely top words include the terms "society", "system", "role", "individual", and "process", we can label this topic "social systems".

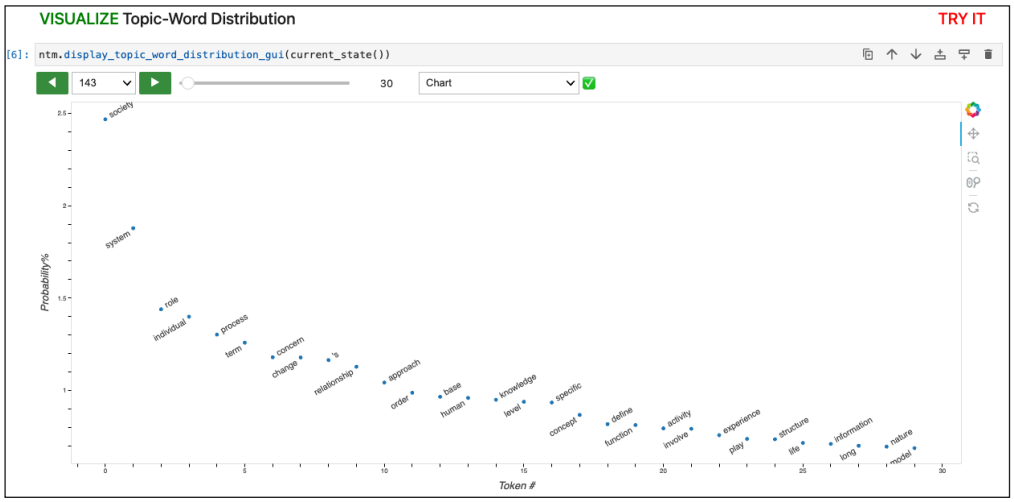


Figure 4 Print screen image of the Notebook feature "Topic-Word-Distribution", focusing on topic 143 (social systems) with its 30 most likely words, each assigned a probability value.

A "topic" generated through LDA topic modeling captures a linguistic structure that resembles what is commonly known as a semantic theme or cluster (Mohr & Bogdanov 2013; Heuser & Le-Khac 2011; Allen & Murdock 2022). It can thus be helpful to study how such a structure behaves over time: when does such a topic emerge, peak, or decrease? If we believe that the topic in question reflects a particular discourse, what can we conclude took place in that discourse at these different junctures? The Notebook feature "Topic Trends over Time" offers one means of addressing these questions: it creates a visualization of the relative strength of a chosen topic (as compared to all other topics in each *Courier* article) by year. Using this tool to chart the trend line of topic 143 produces a graph with striking peaks and valleys (Figure 5), which suggests a range of analytical questions. Why did this discussion of social systems rise in the early 1980s and mid-1990s and then decline thereafter?

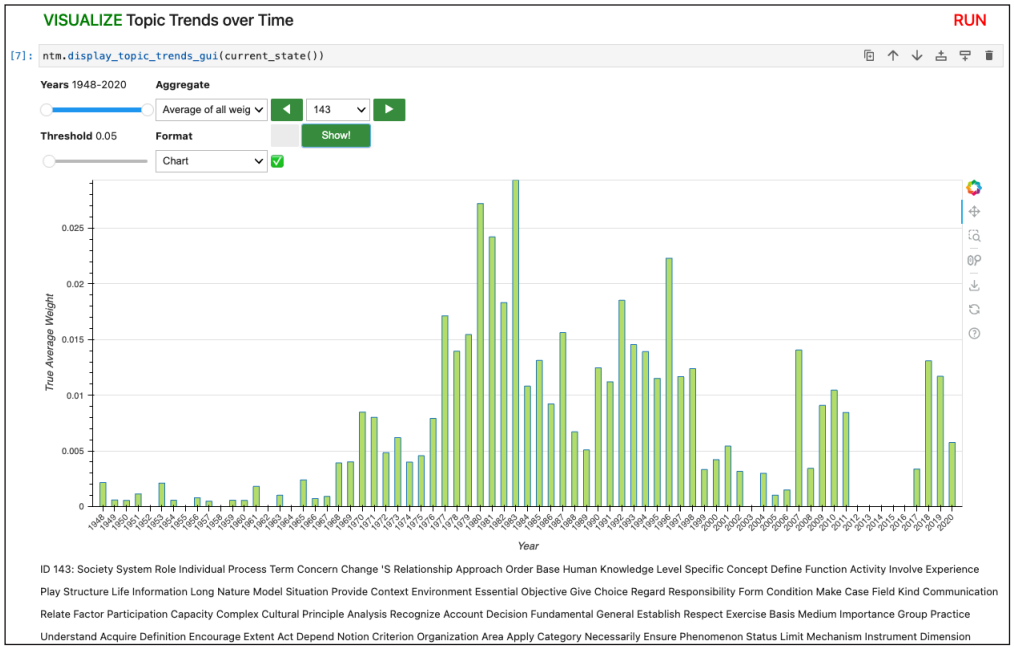


Figure 5 Print screen image of the Notebook feature "Topic Trends over Time", showing the trend line of topic 143 (social systems), using an average weight measure.

To pursue this question, we may want to identify specific articles in *Courier* in which this topic appeared most strongly. Here we deploy the Notebook feature "Browse Topic Documents", which displays the *Courier* articles in which a chosen topic is most probable. Since the LDA model constructs

topics differently than a human reader would, it can point the researcher to topic-related articles of analytical importance that he or she might not have considered otherwise.[13] With this tool, we can use search filters applied to our metadata file to identify articles from a specific time period or to focus only on articles by a named author or on those by UNESCO officials (Figure 6). The tool allows us to read these top articles in full text without leaving the Notebook environment.

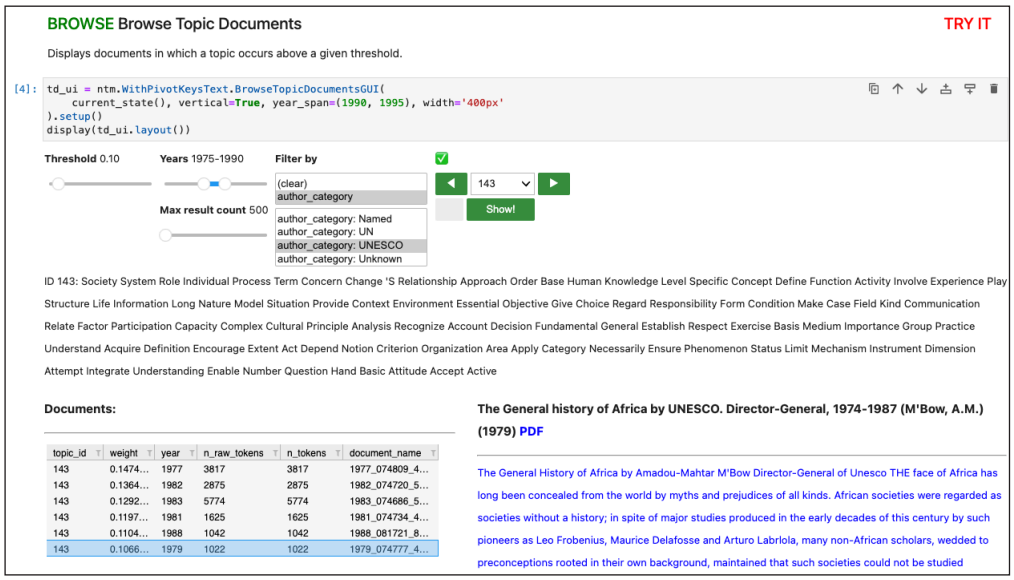


Figure 6 Print screen image of the Notebook feature "Browse Topic Documents", focusing on topic 143 (social systems). Choosing weight threshold 1.0, time period 1975–1990, and only articles that explicitly have "UNESCO" in the author tag, generated six such articles. The article "The General History of Africa" (1979) by former UNESCO Director-General Amadou-Mahtar M'Bow is highlighted, with part of the article text shown to the right, including a link ("PDF") to the corresponding *Courier* issue.

Another exploratory asset of the Jupyter Notebook is its different features for topic network analysis, which enable researchers to study how a topic, such as topic 143, relates to other topics in a given time period. Since the structure of each topic is based on the word distributions in *every* document, it is possible to measure the relationships among those topics that appear (over a certain probability threshold) in the same articles and visualize those relationships as a network. An example of this tool's uses is displayed in Figure 7, in which we have generated a topic network for the period 1975 to 1990.[14] By seeing which topic numbers appear here as linked to topic 143 (social systems) and then checking these numbers on our list of topics, we can see that between 1975 and 1990—the period when topic 143 was at its strongest—the topic was situated in a cluster of topics related to the theme of development. More specifically, we find that topic 143 was directly connected to topic 3 (top words: "international", "programme", "organization", "UNESCO", "country"), topic 5 ("country", "development", "economic", "develop", "world"), topic 45 ("culture", "cultural", "people", "tradition", "civilization"), and topic 50 ("problem", "world", "make", "development", "progress"). This is an indicative result which calls for more research. But it already suggests that when the semantic cluster related to "social systems" was most prevalent in *Courier*, it was invoked in the same context as the themes of development and of UNESCO's own international programs, but also, interestingly, in relation to the theme of culture and tradition.

In the future, the Curated *Courier* 1.0 could be enhanced in several ways. A particular area for extension is the metadata file: tagging the author names by nationality and gender, for example, would make it possible to isolate and analyze articles by writers from certain (groups of) countries or by women and then to compare different subsets of articles divided up in this way. In addition, it might be possible to link many of the authors to other datasets through their Integrated Authority File (GND) identifiers.[15] Linking this data would be particularly interesting, given the wide range of authors who published in *Courier*. In the meantime, the Curated *Courier* 1.0 offers a rich research asset facilitating digital research in the global humanities, whether it is explored with the text analysis methods we have implemented in Jupyter Notebooks or with any number of other tools.

13 This was our experience in conducting research on the concept of nature in *Courier*; see Martin & Mohammadi Norén (2023).

14 The tool's filter options include time period, the topic weight threshold, and the minimum number of documents in which both topics appear. This result can—as for all the Notebook features—be exported for use with other software; in this case, for example, with a network visualization tool like Gephi.

15 See: https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html (accessed 19 January 2024).

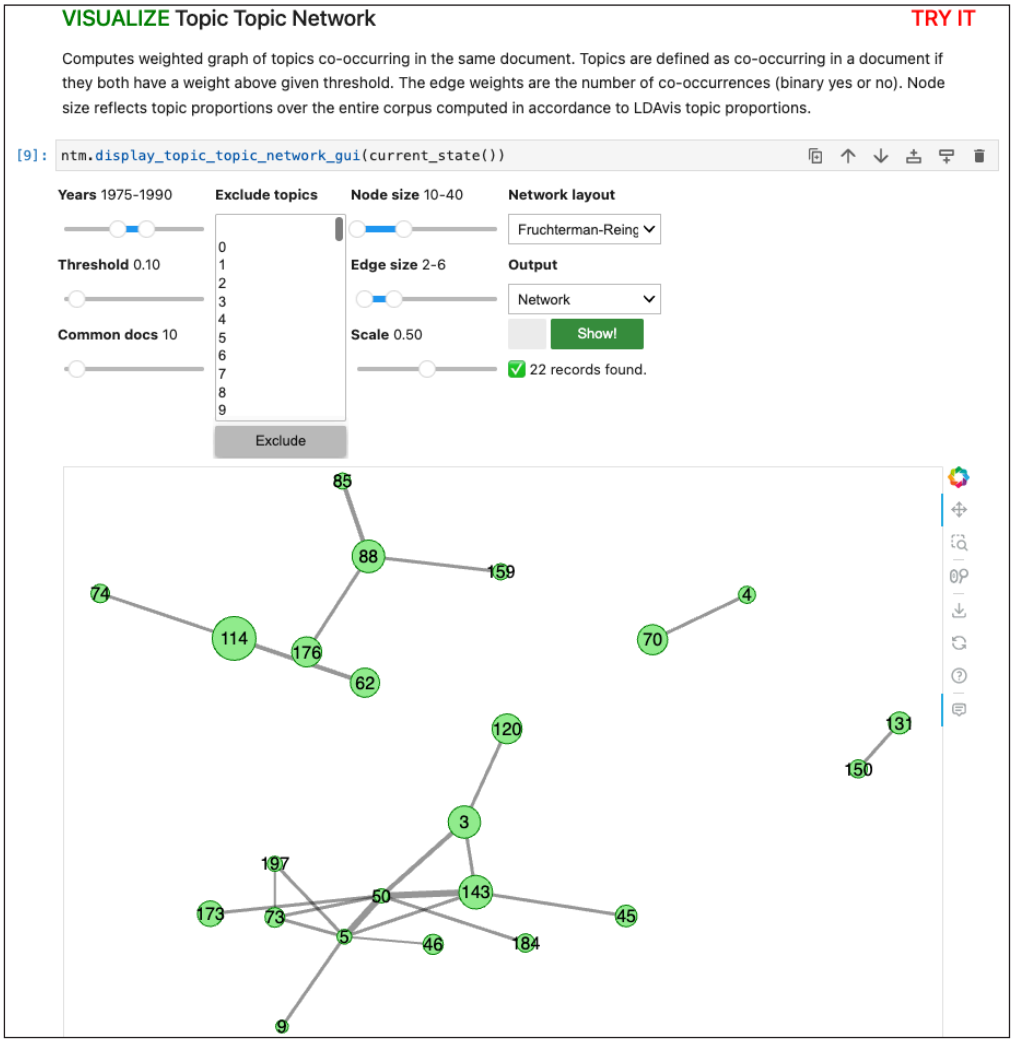


Figure 7 Print screen image of the Notebook feature "Topic Topic Network". Different options are possible for filtering the structure and layout of the topic network. Node labels represent topic numbers.

ACKNOWLEDGEMENTS

Thanks to the UNESCO Library and Archives, and especially archivist Eng Sengsavang, for facilitating access to metadata and documents.

FUNDING INFORMATION

Research for this paper was supported by a grant from the Swedish Research Council (Vetenskapsrådet) for the project "International Ideas at UNESCO: Digital Approaches to Global Conceptual History" (VR 2019-03278).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Benjamin G. Martin orcid.org/0000-0003-3180-2973
Department of History of Science and Ideas, Uppsala University, Uppsala, Sweden

Fredrik Mohammedi Norén orcid.org/0000-0001-8820-1082
Department of Media and Communication Studies, Malmö University, Malmö, Sweden

Roger Mähler orcid.org/0009-0009-4781-9789
Humlab, Umeå University, Umeå, Sweden

Andreas Marklund
Humlab, Umeå University, Umeå, Sweden

Oriane Martin orcid.org/0009-0009-6656-9340
Department of Linguistics, University of Lausanne, Lausanne, Switzerland

REFERENCES

Allen, C., & **Murdock, J.** (2022). LDA Topic Modeling: Contexts for the History and Philosophy of Science. In: G. Ramsey & A. De Block, (Eds.), *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*. pp. 103–119. DOI: https://doi.org/10.2307/j.ctv31djr2f.9

Amirell, S. (2023). From Global Studies to Global Humanities. *Humanities*, *12*(2). DOI: https://doi.org/10.3390/h12020027

Apache PDFBox. (2023). Apache PDFBox – A Java PDF Library. https://pdfbox.apache.org/ [Last accessed 19 January 2024].

Barman, R., Ehrmann, M., Clematide, S., Oliveira, S. A., & **Kaplan, F.** (2021). Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers. *Journal of Data Mining and Digital Humanities*. DOI: https://doi.org/10.46298/jdmdh.6107

Betts, P. (2020). *Ruin and Renewal: Civilising Europe After the Second World War*. Profile Books.

Blei, D. (2012). Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77–84. DOI: https://doi.org/10.1145/2133806.2133826

Blei, D., Ng, A. Y., & **Jordan, M. I.** (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*(1), 993–1022. https://dl.acm.org/doi/10.5555/944919.944937

Brouillette, S. (2019). *UNESCO and the Fate of the Literary*. Stanford, CA: Stanford University Press. DOI: https://doi.org/10.1515/9781503610323

Carruthers, W. (2022). *Flooded Pasts: UNESCO, Nubia, and the Recolonization of Archaeology*. Ithaca, NY: Cornell University Press. DOI: https://doi.org/10.7591/cornell/9781501766442.001.0001

Darian-Smith, E., & **McCarty, P. C.** (2017). *The Global Turn: Theories, Research Designs, and Methods for Global Studies*. Berkeley, CA: University of California Press. DOI: https://doi.org/10.1525/9780520966307

Dausse, A. (2018). Remembering Sandy Koffler, my Grandfather. *The UNESCO Courier*, 4. DOI: https://doi.org/10.18356/b5d181ea-en

Defourny, V. (2003). Public Information in the UNESCO. In: K. Sriramesh & K. Verčič (Eds.), *The Global Public Relations Handbook: Theory, Research, and Practice*. Mahwah, NJ: Lawrence Erlbaum.

Déjean, H., & **Meunier, J.-L.** (2006). A System for Converting PDF Documents into Structured XML Format. In: *Document Analysis Systems VII. DAS 2006. Lecture Notes in Computer Science*, *3872*, 1–12. Berlin: Springer. DOI: https://doi.org/10.1007/11669487_12

Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P., & **Barman, R.** (2020). Language Resources for Historical Newspapers: The Impresso Collection. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 958–968. DOI: https://doi.org/10.5281/zenodo.4641902

Gfeller, A. É., & **Eisenberg, J.** (2016). UNESCO and the Shaping of Global Heritage. In: P. Duedahl (ed.), *A History of UNESCO: Global Actions and Impacts*. pp. 279–99. DOI: https://doi.org/10.1007/978-1-137-58120-4_14

Guldi, J. (2023). *The Dangerous Art of Text Mining: A Methodology for Digital History*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009263016

Hamelink, C. (2012). Global Media Research and Global Ambitions: The Case of UNESCO. In: I. Volkmer (ed.), *The Handbook of Global Media Research*. pp. 28–39. DOI: https://doi.org/10.1002/9781118255278.ch2

Heuser, R., & **Le-Khac, L.** (2011). Learning to Read Data: Bringing out the Humanistic in the Digital Humanities. *Victorian Studies*, *54*(1), 79–86. DOI: https://doi.org/10.2979/victorianstudies.54.1.79

Hurtado Bodell, M., Magnusson, M., & **Mützel, S.** (2022). From Documents to Data: A Framework for Total Corpus Quality. *Socius: Sociological Research for a Dynamic World*, *8*, 1–15. DOI: https://doi.org/10.1177/23780231221135523

Iriye, A. (2002). *Global Community: The Role of International Organizations in the Making of the Contemporary World*. Berkeley: University of California Press. DOI: https://doi.org/10.1525/9780520936126

Jarlbrink, J., & **Snickars, P.** (2017). Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive. *Journal of Documentation*, *73*(6). DOI: https://doi.org/10.1108/JD-09-2016-0106

Kott, S. (2011). International Organizations: A Field of Research for Global History. *Studies in Contemporary History*, *8*(3), 446–450. DOI: https://doi.org/10.14765/zzf.dok-1641

Krebs, E. (2016). Popularizing Anthropology, Combating Racism: Alfred Métraux at The UNESCO Courier. In: P. Duedahl (ed.), *A History of UNESCO: Global Actions and Impacts*. pp. 29–48. DOI: https://doi.org/10.1007/978-1-137-58120-4_2

Mäkelä, E. (2020). Wrangling with Non-Standard Data. In *DHN2020: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference Riga*, Latvia, October 21–23, 2020. CEUR Workshop Proceedings, 2612. https://ceur-ws.org/Vol-2612/paper6.pdf

Martin, B., & **Mohammadi Norén, F.** (2023). Nature and Culture in the Age of Environmental Crisis: Digital Analysis of a Global Debate in The UNESCO Courier, 1948–2020. In: *DHNB 2023 Conference Proceedings: Digital Humanities in the Nordic and Baltic Countries Publications*, *5*(1), 274–86. DOI: https://doi.org/10.5617/dhnbpub.10671

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. https://mallet.cs.umass.edu/diagnostics.php [Last accessed 19 January 2024].

McDonald, P. (2017). *Artefacts of Writing: Ideas of the State and Communities of Letters from Matthew Arnold to Xu Bing*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198725152.001.0001

Meskell, L. (2018). *A Future in Ruins: UNESCO, World Heritage, and the Dream of Peace*. Oxford: Oxford University Press.

Mohr, J., & **Bogdanov, P.** (2013). Introduction—Topic Models: What They Are and Why They Matter. *Poetics*, *41*(6), 545–569. DOI: https://doi.org/10.1016/j.poetic.2013.10.001

Moretti, F. (2013). *Distant Reading*. London: Verso.

Moretti, F., & **Pestre, D.** (2015). Bankspeak. *New Left Review*, *2*(92), 75–99.

Mueller, M., & **Burns, P. R.** (2016). Collaborative Curation and Exploration of the EEBO-TCP Corpus. *New Technologies in Medieval and Renaissance Studies*, *6*, 147–167.

Ren, F., & **Matsumoto, K.** (2016). Semi-Automatic Creation of Youth Slang Corpus and Its Application to Affective Computing. *IEEE Transactions on Affective Computing*, *7*(2), 176–189. DOI: https://doi.org/10.1109/TAFFC.2015.2457915

Simonsen, M. (2020). Routes of Knowledge: The Transformation and Circulation of Knowledge in the UNESCO Courier, 1947–1955. In: J. Östling, D. Larsson Heidenblad & A. Nilsson Hammar (Eds.), *Forms of Knowledge: Developing the History of Knowledge*. pp. 225–240. DOI: https://doi.org/10.2307/jj.919496.17

Smith, R. (2007). An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*. DOI: https://doi.org/10.1109/ICDAR.2007.4376991

Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226612973.001.0001

Voormann, H., & **Gut, U.** (2008). Agile Corpus Creation. *Corpus Linguistics and Linguistic Theory*, 4(2). DOI: https://doi.org/10.1515/CLLT.2008.010

Wong, L. E. (2008). Relocating East and West: UNESCO's Major Project on the Mutual Appreciation of Eastern and Western Cultural Values. *Journal of World History*, *19*(3), 349–374. DOI: https://doi.org/10.1353/jwh.0.0019

TO CITE THIS ARTICLE:
Martin, B. G., Norén, F. M., Mähler, R., Marklund, A., & Martin, O. (2024). The Curated *UNESCO Courier* 1.0: Annotated Corpora for Digital Research in the Global Humanities. *Journal of Open Humanities Data*, 10: 20, pp. 1–13. DOI: https://doi.org/10.5334/johd.181

Submitted: 09 November 2023
Accepted: 19 January 2024
Published: 21 February 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.