Journal of Open Humanities Data

Anderson BG 2020 The Cybernetics Thought Collective: Machine-Generated Data Using Computational Methods. *Journal of Open Humanities Data,* 6: 7. DOI: https://doi.org/10.5334/johd.19

DATA PAPER

# The Cybernetics Thought Collective: Machine-Generated Data Using Computational Methods

Bethany G. Anderson
University of Illinois at Urbana-Champaign, US
bgandrsn@illinois.edu

This dataset comprises machine-generated data from the research records and personal archives of four founding members of the transdisciplinary field of cybernetics—W. Ross Ashby, Warren S. McCulloch, Heinz von Foerster, and Norbert Wiener. These archives (or, *fonds*) are held by the British Library, the American Philosophical Society, the University of Illinois at Urbana-Champaign, and MIT, respectively. The data were created for "The Cybernetics Thought Collective: A History of Science and Technology Portal Project" (2017–2019), a pilot project funded by the National Endowment for the Humanities (NEH). Using computational methods and tools—machine learning, named entity recognition, and natural language processing—on digitized archival records, the data were generated to enhance archival access in three distinct but interrelated ways: as archival metadata for the digitized records, as reusable data to facilitate digital scholarly analyses, and as the basis for a series of test visualizations. The data represent entities associated with cybernetic concepts and the main actors attached to the cybernetics movement and the exchange of its ideas. The dataset is stored along with the digitized records in the University of Illinois (U of I) Library's multi-tiered repository, a replicated preservation service based on PREMIS (Preservation Metadata: Implementation Strategies). Reuse potential for this dataset includes historical/archival, linguistic, and artistic analyses of the data to examine connections between the cybernetic entities.

**Keywords:** Archive records; Science and technology; Social networks
**Funding statement:** "The Cybernetics Thought Collective: A History of Science and Technology Portal Project" (NEH PW-253912-17), was funded by the National Endowment for the Humanities' Humanities Collections and Reference Resources program (US).

## (1) Overview

### Repository location

https://digital.library.illinois.edu/collections/38ec6eb0-18c3-0135-242c-0050569601ca-1.

### Context

Cybernetics was a transdisciplinary scientific movement in the mid-twentieth century that emerged from the Macy Conferences on "Circular Causal and Feedback Mechanisms in Biological and Social Systems" (1946–1953), as well as the publication of Norbert Wiener's *Cybernetics: Or the Control and Communication in the Animal and the Machine* [34]. As a postwar movement that explored the possibilities and implications of "thinking machines" amid broader social currents, cybernetics inspired questions about what it means to be a human being, a machine, or a social system. Discussions radiating from the Macy Conferences evolved into correspondence networks, publications, and the establishment of centers for systems and cybernetics.

The data were created for "The Cybernetics Thought Collective: A History of Science and Technology Portal Project," a grant project funded by the National Endowment for the Humanities' Humanities Collections and Reference Resources program (US) [29]. The project was led by the University of Illinois at Urbana-Champaign Library in collaboration with the American Philosophical Society, the British Library, and the MIT Distinctive Collections. By making the data available, the project seeks to reveal insights about the cybernetics phenomenon through the "thought collective" [11] that exchanged and interrogated ideas through correspondence and other records.

## (2) Methods

### Steps

#### *Digitization and OCR*

The four participating institutions identified correspondence, scientific journals, and publications from the personal archives (or *fonds*) of W. Ross Ashby, Warren S.

McCulloch, Heinz von Foerster, and Norbert Wiener for digitization. In total, 61,067 pages of archival records were digitized, resulting in 615 digital objects (which represent folder-level or multi-page item-level aggregations of digitized records). The project created PDFs for archival access purposes as well as high-resolution preservation TIFF files. The former were processed by optical character recognition (OCR) software to make the records machine-readable. Some materials are also handwritten and were transcribed as time allowed.

### *Normalization and Input Creation*

PDFMiner [20] was used to extract text from the OCR-ed records into plaintext files. Before testing entity extraction, natural language processing, and machine learning software, text remediation and normalization was needed to both address OCR errors and to translate some of the *fonds*' Italian, Spanish, French, and German texts into English. Translation was completed with the aid of N-grams and Googletrans [17], while Wolfram Text Analysis tools [35] were used to remove stopwords.

Concurrent with this step, the project team created inputs, or a cybernetics vocabulary. The project sought to specifically identify and extract cybernetic entities; fortunately, cybernetics has a distinct set of core concepts related to behavior, self-organization, and feedback mechanisms, from which a vocabulary could be derived [2, 25]. Identifying this vocabulary was especially important for connecting concepts and agents to each other in the cybernetics network. The project team used *Cybernetics of Cybernetics: Or, the Control of Control and the Communication of Communication* as a source for generating a cybernetics vocabulary [32]. *Cybernetics of Cybernetics* is a compilation that prominently features Ashby, McCulloch, von Foerster, Wiener, and key cybernetic ideas at the time that they were active in the transdiscipline. A digital version of the text was run through Voyant Tools [28] to generate a list of keywords based on frequency. This list was narrowed to include the most frequently occurring terms (about 200 total). Members of the project's advisory board (who comprised technologists and subject-experts in cybernetics) reviewed this list and offered additional suggestions.

### *Entity Extraction, Natural Language Processing, and Classification*

Using this cybernetic vocabulary as inputs, one of the project's programmers experimented with a number of Python libraries for natural language processing and named entity extraction (e.g., NLTK [6] and spaCy [24]) and the University of Illinois Cognitive Computation Group's NLP pipeline software [7]. Following entity identification and extraction, the project team decided to adopt a supervised machine learning approach to classify the records into four broad categories: Mathematics/Logic, Computers/Machines, Psychology/Neuroscience, and Personal. Naïve Bayes [8] and Weka [14] were used for the machine learning portion of the project. Percentages of certainty for the classifications were also generated through this process. Additional testing was performed with sentiment analysis using NLTK and VADER [19].

### *Text Processing and Remediation Pipeline*

After testing various software, a Python-based pipeline was developed (see **Figure 1**). Following the pipeline, the project team imported text from the PDF files into plaintext; normalized the files; removed files that contained a significant amount of noise that could not be easily remediated with existing tools in the allotted timeframe; identified the language of the documents and translated into English where necessary; extracted entities; classified the documents into categories; and estimated the percentage of certainty for each category per document. The pipeline is documented in the project's GitHub Repository [30]. All of the data resulting from the entity extraction, sentiment analysis, and machine learning steps were imported into a CSV file that was made available for research use and used as metadata for the digital collection. A more detailed overview of the methodology and the steps employed is delineated in the project's white paper [4]. It is important to note that the creation of this pipeline was not a linear process and involved retesting tools and revisiting several steps.

### *Preservation and Access*

The PDFs of the digital surrogates and files containing the inputs, classification data, percentages of certainty for those classifications, and extracted entities were ingested into the University of Illinois Library's digital repository service for preservation and access. The digital repository (known as Medusa [31]) is a replicated multi-tiered Fedora-based repository that uses PREMIS [26]. The classification data, percentages of certainty, and entities also populate the metadata application profile for each PDF in the repository's access interface. The data were made available as a dataset via a CSV file for users to download, along with a CSV file containing the original inputs and a readme file that provides additional information about the data and the process that created them [3]. The dataset includes file-level metadata, some of which is human created (e.g., level of description and title), and some of which is collection-level metadata that applies to all digital objects in the same *fonds* (e.g., scope and contents, parent collection, collection identifier), and provides original archival context for the machine-generated data (e.g., machine-extracted feature, cybernetic classification, certainty). These fields are described more fully in the data dictionary in the readme file. A selection of the data was also used to create test visualizations, which are available on the project site.

## Sampling strategy

This pilot project aimed to produce a proof-of-concept machine learning, named entity recognition, and natural language processing pipeline for meta/data generation and classification of archival records; through this process, a representative sample of documents that illustrate prominent cybernetic concepts and consist of letters between von Foerster, Ashby, McCulloch, Wiener, and other known cyberneticians were selected from across the four *fonds*. However, statistical sampling techniques were employed at various stages of the natural language processing and

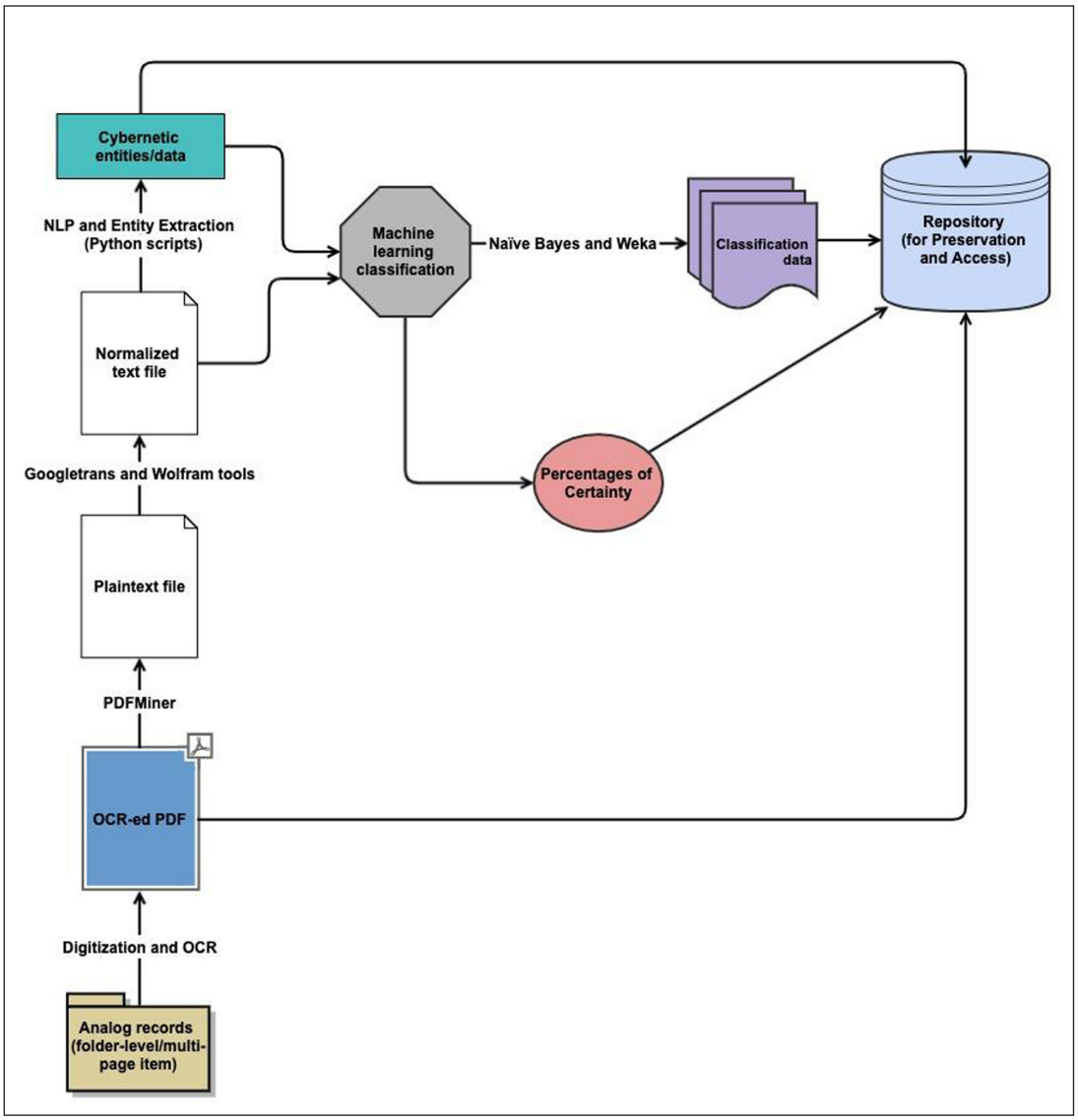


**Figure 1:** Illustration of extraction, classification, and preservation/access pipeline.

entity-extraction workflow. For example, to translate texts into English, a test set of approximately 200 documents in English, German, French, and Italian was created in order to employ an N-gram approach to language identification. The Python library Googletrans was then used to translate the texts into English. Additionally, a training set of 154 documents from all *fonds* were manually annotated and prepared for the supervised classification model.

### Quality Control

The majority of the records from the four *fonds* are typewritten; these records were processed with OCR software, and, as time allowed, handwritten documents were transcribed. The texts also required "normalization" in order to be machine-ready. After extracting the text from the OCR-ed records to import into plaintext files, character errors that resulted from OCR were remediated (e.g., extra spaces between letters in a word, or alpha-numeric characters that were misread as non-ASCII characters).

Statistical analysis was performed on the extracted entities to identify which entities surface the most frequently in the corpus, as a means of determining which entities appear most significant. We tested this through N-grams and Term-Frequency Inverse Document Frequency (TF-IDF) to determine the frequency of an entity in each document and thus its importance throughout the entire corpus. Using TF-IDF in an archival context has precedent ([10], pp. 109–110), so we hoped that it would have utility for the project. The team felt this would be useful for comparison against the original cybernetic inputs. However, despite removing "noise" such as stop-words (i.e., commonly used words like "the," "of," or "but"), TF-IDF proved not as reliable as an N-gram approach for determining entity relevancy within the corpus. For TF-IDF to produce more useful results, document-length would need to be normalized. Given the overall nonuniformity of archival records in this particular corpus (and in archival *fonds* in general), it is difficult to normalize records for length.

To assess the accuracy of the machine learning results, the project team used Weka to perform a chi-squared analysis to help us better understand the accuracy of the training set in the classification process. The results revealed 71.1% "true positives" and 4% "false positives," indicating that the majority of the entities were useful in informing

which documents were classified into specific categories. However, a manual analysis revealed more false positives (i.e., a few inaccurately classified documents). This assessment enabled the project team to perform a degree of quality control on the dataset and understand how we might improve the machine learning results in the future, especially by creating a larger training set.

### Assessment

As a proof of concept, the Cybernetics Thought Collective project opened up the possibility of applying computational methods to archival records. But it also opened up questions about how to develop and streamline computational workflows in an archival setting, how best to document those workflows to facilitate data reuse and reproducibility, and to provide transparency so that users can understand the "computational provenance" of the results.

While the results of the project did reveal connections between documents across the four *fonds* through the extracted entities, the machine learning results indicated a need for additional refinement. For example, some of the documents which almost exclusively consisted of discussions of a technical nature were classified as "Personal." Thus, we will need larger training sets in tandem with better quality control mechanisms to produce more reliable results. Participants in computational archival projects need to be able to anticipate the labor necessary for creating viable inputs and training sets and for verifying the trustworthiness of the results.

Computational archival projects require close collaboration between archivists, programmers, data curators, and digital preservationists, who each provide vital input and expertise at different decision points. Likewise, in the future more engagement with potential users will be vital for determining the utility of the results and their implications for archival research, which should also inform the creation and refining of processes that generate these datasets.

The project raised questions about the relationship between machine-generated archival datasets and the original archival records—especially how that relationship is represented in both archival systems and visualization interfaces in order to ensure the original "archival provenance" of the data and materials from which they are derived are clearly described to prevent decontextualization. Digital records, and the data generated from them, can provide greater context and enhance access to each other. Therefore, it is important to find ways to make them mutually discoverable in archival access systems.

## (3) Dataset description

### Object name

CTC_Machine-Generated-Data.csv

### Format names and versions

CSV

### Creation dates

2017-10-25 to 2018-05-07

### Dataset Creators

1. Bethany Anderson, University of Illinois at Urbana-Champaign (conceptualization, data curation, funding acquisition, methodology; project administration, resources, supervision)
2. Christopher J. Prom, University of Illinois at Urbana-Champaign (conceptualization, data curation, funding acquisition, methodology, project administration, resources)
3. Anirudh Chandrashekhar (data curation, formal analysis, investigation, methodology, software, validation)
4. Saumye Kaushik (data curation, formal analysis, investigation, software)
5. Alex Dolski, University of Illinois at Urbana-Champaign (investigation, methodology, software)
6. James A. Hutchinson, University of Illinois at Urbana-Champaign (conceptualization, data curation, funding acquisition)
7. Mark Sammons, University of Illinois at Urbana-Champaign (methodology, resources, software)
8. Kevin Hamilton, University of Illinois at Urbana-Champaign (conceptualization, funding acquisition)
9. Charles Greifenstein, American Philosophical Society (funding acquisition, methodology, resources)
10. Jonathan Pledge, British Library (funding acquisition, methodology, resources)
11. Thomas Rosko and Beverly Turner, MIT (funding acquisition, methodology, resources)

### Language

English, French, Italian, German, and Spanish.

### License

CC BY 4.0

### Repository name

University of Illinois Digital Collections repository

### Publication date

2020-01-24

## (4) Reuse potential

While cybernetics experienced a heyday that spanned the mid-late twentieth century, its philosophical influences are widespread. Indeed, vestiges of cybernetics continue to surface in modern computing, information theory, and cognitive science, as just a few examples [21, 9]. Because of this intellectual omnipresence, the data have the potential to shed light on the etymology of concepts and disciplinary areas of specialization. For example, the data may be useful for contributing to discussions about artificial intelligence and its relationship to cybernetics [12, 27]. However, these data do not provide insight into the evolution of the terms themselves or how the relationships between entities shifted and changed over time.

From a historical perspective, the data can be reused to reveal additional connections between cybernetic entities and the scientists who formed the cybernetics movement.

Cybernetics continues to be of recent interest to historians and science and technology studies scholars (for example [21, 1]). Since this was a pilot project that resulted (in part) in several test visualizations, the data are also available for bulk download to facilitate use in other digital scholarship projects. We hope that this opens the data up to new questions and explorations of the boundaries of the "thought collective," while also serving as a step toward meeting emergent research needs within a digital scholarship framework (for example [18]).

An important aspect of data reuse is providing sufficient contextual information to enable a variety of reuse(s). Because the dataset includes information about the original digital records from which the entities are generated, and thus the original *fonds*, this may lead to new pathways to the digitized records themselves that are in line with FAIR data reuse principles for archival materials [22]. At the same time, the relative success with which researchers are able to reuse the data and gain new insights can inform the project's future phases as it refines its software pipeline and methods for assessing quality control. It is hoped that this data paper provides additional information about the process that generated the data, so that others may test its reproducibility and assess the results.

It is worth noting that reuse should also logically extend to the digital records themselves; all digitized materials have been made machine-readable and are accessible through the University of Illinois' repository/digital collections portal. Users can download the OCR-ed records, process them through different software pipelines, and perform their own computational analyses. While the methods employed by this project sought to extract data from the records, drawing a distinction between the reuse potential of the records themselves and the data generated from them is somewhat blurry given the interdependence of the data on the records to elucidate their context(s) and make them reusable [15]. It is thus important to emphasize that the digital records themselves are (re)usable. A future phase of this project will seek to engage researchers and the archival community in identifying additional reuse cases for both the data and the digitized records themselves, and investigate the possibility of interactive interfaces that open up explorations of records as data and user-driven reorderings of content [23, 36].

The data also have potential reuse value in a visual culture space. Cybernetics (especially second-order cybernetics) invoked visual and art historical references to interrogate and illustrate many of its ideas. For example, to peruse the publications that emerged from the Biological Computer Laboratory—the center for cybernetics at the University of Illinois directed by Heinz von Foerster—is to become simultaneously immersed in scientific diagrams and esoteric imagery of ouroboros and art historical iconography (see, for example [33]). Cybernetics has inspired "cybernetic art" and explorations of media culture through a cybernetic lens [5, 13]. Examples of cybernetic data either informing or becoming artistic works themselves also have precedent, indicating that such reuses are not unimaginable [16]. The data resulting from the project can contribute to cybernetic explorations at the intersection of art, technology, and new media.

## Additional File

The additional file for this article can be found as follows:

- **Readme for the Cybernetics Thought Collective Data.** This readme file contains a brief description of the dataset, metadata fields, and the process of data creation. https://digital.library.illinois.edu/items/3cd33c50-8c95-0138-729a-02d0d7bfd6e4-8.

## Acknowledgements

Thank you to the National Endowment for the Humanities for providing funding that supported this project. I would like to express my deep gratitude to my collaborators on this project—Christopher J. Prom, James A. Hutchinson, Kevin Hamilton, Alex Dolski, Mark Sammons, Charles Greifenstein, Jonathan Pledge, Thomas Rosko, and Beverly Turner. I am also grateful for the project's advisory board members who generously shared their time and insights. Special thanks to Stephen Wolfram and Jesús V. Hernández of Wolfram Research for donating time and technology resources to the project. I especially want to thank Anirudh Chandrashekhar, whose work was crucial to the project's success. Many colleagues offered advice and guidance during the project, especially William J. Maher, MJ Han, Patricia Lampron, Angela Waarala, Tom Habing, Kyle Rimkus, Jennifer Hain Teper, and Kathie Veach. In addition, there were many other contributors to the project: Christine Pallon, Shreya Udhani, Brinna Michael, Alicia Hopkins, Tanairy Delgado, Saumye Kaushik, and Meghna Shrivastava. Lastly, many thanks to Heidi Imker and Kelli Trei for providing invaluable feedback on this paper.

## Competing Interests

The author has no competing interests to declare.

## Author Contributions

Conceptualization; Data curation; Funding acquisition; Methodology; Project administration; Supervision; Writing – original draft.

## References

1. **Abraham TH.** *Rebel Genius: Warren S. McCulloch's Transdisciplinary Life in Science.* The MIT Press. 2016. DOI: https://doi.org/10.7551/mitpress/9780262035095.001.0001.
2. **American Society for Cybernetics.** (n.d.). *ASC Glossary*. Retrieved May 28, 2020, from http://www.asc-cybernetics.org/foundations/ASCGlossary.htm.
3. **Anderson BG.** *Data from the Cybernetics Thought Collective* [Data set]. University of Illinois Digital Library. 2020. https://digital.library.illinois.edu/items/3c80ad40-8c95-0138-729a-02d0d7bfd6e4-b.
4. **Anderson BG, Prom CJ, Hutchinson JA, Chandrashekhar A, Michael B, Udhani S, Sammons M, Dolski A, Hamilton K, Kaushik S, Shrivastava M.** *The Cybernetics Thought Collective: A History of Science and Technology Portal Project* [White paper]. University of Illinois at Urbana-Champaign. 2019. https://www.ideals.illinois.edu/handle/2142/106050.

5. **Archive of Digital Art.** n.d. *Roy Ascott*. Retrieved May 28, 2020, from https://www.digitalartarchive.at/database/artists/general/artist/ascott.html.
6. **Bird S, Loper E, Klein E.** *NLTK* (Version 3.2.5) [Computer software]. NLTK Documentation. 2017. https://www.nltk.org/.
7. **Cognitive Computation Group.** *CogComp NLP Pipeline* [Computer software]. GitHub. 2017. https://github.com/CogComp/cogcomp-nlp/tree/master/pipeline.
8. **Dawson R.** *Bayesian Classifier* [Computer software]. GitHub. 2016. https://github.com/codebox/bayesian-classifier.
9. **Dupuy J.** *On the Origins of Cognitive Science: The Mechanization of the Mind*. 2009. The MIT Press.
10. **Esteva M.** *The Aleph in the Archive: Appraisal and Preservation of a Natural Electronic Archive* [Doctoral dissertation, University of Texas at Austin]. Texas ScholarWorks. 2008. https://repositories.lib.utexas.edu/handle/2152/3840.
11. **Fleck L.** *Genesis and Development of a Scientific Fact*. Trenn TJ, Merton RK (eds.). 1979. University of Chicago Press. (Original work published 1935).
12. **Franchi S, Guezeldere G, Minch E.** Interview with Heinz von Foerster. *Stanford Humanities Review*. 1995; 4(2): 288–307.
13. **Fuller M.** *Media Ecologies: Materialist Energies in Art and Technoculture*. 2007. The MIT Press.
14. **Frank E, Hall MA, Witten IH.** *Weka* (Version 3.8) [Computer software]. University of Waikato. 2016. https://www.cs.waikato.ac.nz/ml/weka/.
15. **Grant R.** Recordkeeping and Research Data Management: A Review of Perspectives. *Records Management Journal*. 2017; 27(2): 159–174. DOI: https://doi.org/10.1108/RMJ-10-2016-0036.
16. **Hamilton K.** (n.d.). *BCL/IGB Mural*. KevinHamilton.org. http://www.kevinhamilton.org/bcl_igb/.
17. **Han S.** *Googletrans* (Version 2.1.4) [Computer software]. Python Package Index. 2017. https://pypi.org/project/googletrans/.
18. **Harris G, Potter A, Zwaard K.** *Digital Scholarship at the Library of Congress*. Library of Congress. 2020. https://labs.loc.gov/static/labs/work/reports/DHWorkingGroupPaper-v1.0.pdf.
19. **Hutto CJ, Gilbert EE.** *VADER Sentiment Analysis* [Computer software]. GitHub. 2014. https://github.com/cjhutto/vaderSentiment.
20. **Jeong J.** *PDFMiner* [Computer software]. GitHub. 2016. https://github.com/jaepil/pdfminer3k.
21. **Kline RR.** *The Cybernetics Moment: Or Why We Call our Age the Information Age*. 2015. The MIT Press.
22. **Koster L, Woutersen-Windhouwer S.** FAIR Principles for Library, Archive and Museum Collections: A Proposal for Standards for Reusable Collections. *Code4Lib*. 2018; 48. https://journal.code4lib.org/articles/13427.
23. **Lemieux VL.** Toward a 'Third-Order' Archival Interface: Research Notes on Some Theoretical and Practical Implications of Visual Explorations in the Canadian Context of Financial Electronic Records. *Archivaria*. 2014; 78: 53–93. https://archivaria.ca/index.php/archivaria/article/view/13721.
24. **Montani I.** *spaCy* (Version 2.0) [Computer software]. GitHub. 2017. https://github.com/explosion/spaCy.
25. **Principia Cybernetica Web.** *Web Dictionary of Cybernetics and Systems*. 2002. Retrieved May 28, 2020, from http://pespmc1.vub.ac.be/ASC/INDEXASC.html.
26. **Rimkus K, Habing T.** Medusa at the University of Illinois at Urbana-Champaign: A Digital Preservation Service Based on PREMIS. *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. 2013. DOI: https://doi.org/10.1145/2467696.2467725.
27. **Sato K.** From AI to Cybernetics. *AI &* Society. 1991; 5: 155–161. DOI: https://doi.org/10.1007/BF01891721.
28. **Sinclair S, Rockwell G.** *Voyant Tools* [Computer software]. 2020. https://voyant-tools.org/.
29. **University of Illinois Archives.** *Cybernetics Thought Collective Project*. 2019. Retrieved May 27, 2020, from https://archives.library.illinois.edu/thought-collective/.
30. **University of Illinois Archives.** *cybernetics-thought-collective* [Computer software]. GitHub. 2018. https://github.com/cybernetics-thought-collective.
31. **University of Illinois Library.** *Medusa*. 2020. Retrieved October 8, 2020, from https://medusa.library.illinois.edu/static_pages/technology.
32. **Von Foerster H.** (Ed.). *Cybernetics of Cybernetics, or the Control of Control and the Communication of Communication*. 1974. Biological Computer Laboratory.
33. **Von Foerster H.** *On Constructing a Reality* (Report No. 234). BCL Publication, University of Illinois at Urbana-Champaign. 1973. https://digital.library.illinois.edu/items/3f260d50-29ac-0136-4d81-0050569601ca-0.
34. **Wiener N.** *Cybernetics: Or, Control and Communication in the Animal and the Machine*. J. Wiley. 1948.
35. **Wolfram** (n.d.). *Text Analysis* [Computer software]. Wolfram Language and System Documentation Center. https://reference.wolfram.com/language/guide/TextAnalysis.html.
36. **Yeo G.** Bringing Things Together: Aggregate Records in a Digital Age. *Archivaria*. 2012; 74: 43–19. https://archivaria.ca/index.php/archivaria/article/view/13407.


**How to cite this article:** Anderson BG 2020 The Cybernetics Thought Collective: Machine-Generated Data Using Computational Methods. *Journal of Open Humanities Data,* 6: 7. DOI: https://doi.org/10.5334/johd.19

**Published:** 27 October 2020



]u[ *Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS