



# When Text and Speech are Not Enough: A Multimodal Dataset of Collaboration in a Situated Task

DATA PAPER

IBRAHIM KHEBOUR 

RICHARD BRUTTI 

INDRANI DEY 

RACHEL DICKLER 

KELSEY SIKES 

KENNETH LAI 

MARIAH BRADFORD 

BRITTANY CATES 

PAIGE HANSEN 

CHANGSOO JUNG 

BRETT WISNIEWSKI 

CORBYN TERPSTRA 

LEANNE HIRSHFIELD 

SADHANA PUNTAMBEKAR 

NATHANIEL BLANCHARD 

JAMES PUSTEJOVSKY 

NIKHIL KRISHNASWAMY 

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

## ABSTRACT

To adequately model information exchanged in real human-human interactions, considering speech or text alone leaves out many critical modalities. The channels contributing to the “making of sense” in human-human interactions include but are not limited to gesture, speech, user-interaction modeling, gaze, joint attention, and involvement/engagement, all of which need to be adequately modeled to automatically extract correct and meaningful information. In this paper, we present a multimodal dataset of a novel situated and shared collaborative task, with the above channels annotated to encode these different aspects of the situated and embodied involvement of the participants in the joint activity.

]u ubiquity press

## CORRESPONDING AUTHOR:

Nikhil Krishnaswamy

Department of Computer  
Science, Colorado State  
University, Fort Collins, CO, USA  
[nkrishna@colostate.edu](mailto:nkrishna@colostate.edu)

## KEYWORDS:

multimodal interaction;  
collaboration; problem solving;  
situated tasks

## TO CITE THIS ARTICLE:

Khebour, I., Brutti, R., Dey, I., Dickler, R., Sikes, K., Lai, K., Bradford, M., Cates, B., Hansen, P., Jung, C., Wisniewski, B., Terpstra, C., Hirshfield, L., Puntambekar, S., Blanchard, N., Pustejovsky, J., & Krishnaswamy, N. (2024). When Text and Speech are Not Enough: A Multimodal Dataset of Collaboration in a Situated Task. *Journal of Open Humanities Data*, 10: 7, pp. 1–7. DOI: <https://doi.org/10.5334/johd.168>

# 1 OVERVIEW

**Repository location** Zenodo: <https://zenodo.org/records/10252341>

Khebour et al.  
Journal of Open  
Humanities Data  
DOI: 10.5334/johd.168

2

## CONTEXT

Making sense of face-to-face human-human interaction routinely involves more than just spoken or written language; other modalities include but are not limited to gesture, user-interaction modeling, gaze, joint attention, and markers of involvement/engagement. This is particularly true in the case of modeling human *collaboration*. For instance, having students engage in collaborative problem solving (CPS) has been shown to be an effective pedagogical technique that is correlated with positive learning outcomes (Gillies, 2008; Langer-Osuna Gargroetzi, Munson, & Chavez, 2020; Roschelle & Teasley, 1995), and linguistic discourse alone does *not* reliably indicate effective collaboration.

The Weights Task Dataset (WTD) is a novel dataset of a situated, shared collaborative task, originally collected to study multimodal indicators of collaborative problem solving. This dataset complements other datasets for human-human interaction such as Anderson et al. (1991); Liu, Cai, Ji, and Liu (2017); Van Gemeren, Poppe, and Veltkamp (2016); Wang et al. (2017); Yun, Honorio, Chattopadhyay, Berg, and Samaras (2012), which lack at least one of: multimodal data, physical object manipulation, or multiparty interaction. Our data is novel in the joint presence of speech, gestures, and actions in a collaborative *multiparty* task. Annotation encodes many cross-cutting aspects of the situated and embodied involvement of the participants in joint activity.

## 2 METHOD

The Weights Task is completed by triads at a round table. A webcam captures the task equipment and participants. Kinect Azure cameras capture RGBD video from different angles. Task equipment includes 6 blocks (of varying weight, size, and color), a balance scale, a worksheet, and a computer with a survey where participants submit their answers.

### 2.1 STEPS

Participants (English speakers,  $\geq 18$  years) were recruited from the student body of Colorado State University. Informed consent was obtained. Table 1 shows breakdown of gender and ethnicity.

MALE	FEMALE	CAUCASIAN NON-HISPANIC	HISPANIC/ LATINO	ASIAN
80%	20%	60%	10%	30%

Participants are given a balance scale to determine the weights of five blocks. They are given the weight of one of the blocks (10g), and must determine the weights of the others. As the weight of each block is discovered, it is placed on the worksheet in the cell corresponding to the weight. Next, participants are given a new block and must identify its weight without the scale, by deducing it based on the pattern observed in the initial block weights. Finally, participants must infer the weight of the next hypothetical block in the set and explain how they determined it. After each stage, groups submit their answers in the survey form.

The dataset consists of 10 videos (~170 minutes). Table 2 provides descriptive statistics of the data. Figure 1 shows participants engaging with the objects on the table from the perspective of the main Kinect. Figure 2 shows different annotations (described below).

### Utterance Segmentation and Transcription

Audio from all groups were segmented into utterances, or a single person's continuous speech, delimited by silence, and transcribed. Segmentation and transcription were conducted by humans, by Google Cloud ASR (Velikovich et al., 2018), and by OpenAI's Whisper model (Radford et al., 2023). Human transcription was performed by listening and transcribing what was said by each participant during a given manually-segmented utterance. Google and Whisper transcriptions were conducted over the utterances segmented by the same system (which may conflate overlapping speech by multiple people). Transcriptions are presented in .csv files.

**Table 1** Participant pool distribution of gender and ethnic background. The task was conducted in English. Native languages besides English included Assamese, Bengali, Gujarati, Hindi, Malayalam, Persian, Spanish, Telugu, and Urdu.



	AVG.	SD	MIN.	MAX.
Participant age (yrs.)	24.58	4.58	19	35
Video length (mins.)	17.00	7.00	9	34

**Figure 1** Three participants engaged in the Weights Task. Participant #3 (on the right) is taking a block off the scale to try another configuration while Participant #2 (in the middle) wants to clarify the weight of the block under it. Multimodal information is required to make such a judgment.

**Table 2** Dataset descriptive statistics.

### Collaborative Problems Solving (CPS) Facets

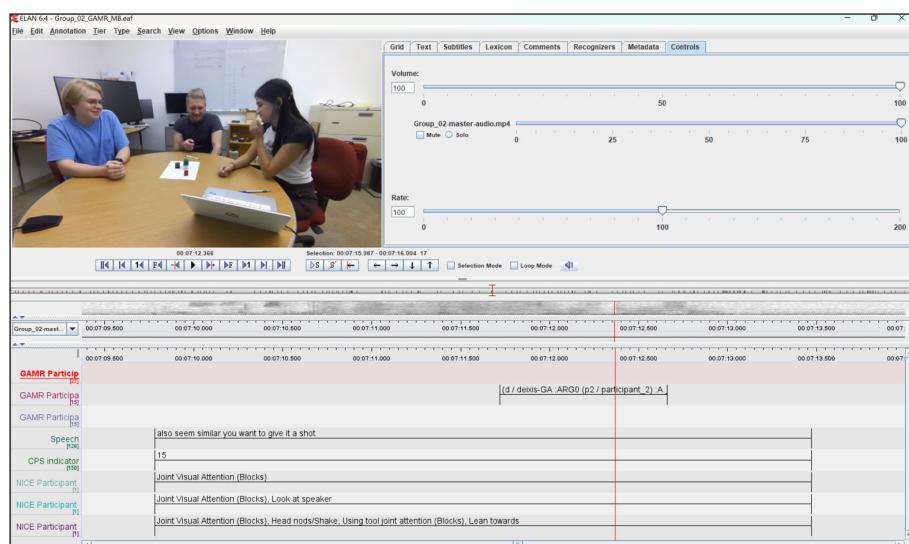
CPS coding is performed at the utterance level using the framework of Sun et al. (2020). Annotators watched the video and coded each utterance with potentially multiple labels based on content, context, and position in the conversational sequence. Videos were annotated by two annotators ( $\kappa = 0.62$ ) and adjudicated by an expert who underwent extensive training in the framework. CPS is presented in .csv files.

### Gesture Abstract Meaning Representation (GAMR)

Participant gestures are annotated using the GAMR framework (Brutti, Donatelli, Lai, & Pustejovsky, 2022). Most WTD gestures are *deictic*, indicating reference to an object or a location. *Iconic* gestures represent attributes of an action or object. The meaning of *emblematic* gestures is set by cultural convention. GAMR was dual annotated by annotators trained by authors of the framework (SMATCH F1-score = 0.75). This data is presented in PENMAN notation in .eaf files.

### Nonverbal Indicators of Collaborative-Learning Environments (NICE)

The NICE coding scheme (Dey et al., 2023) captures nonverbal behaviors when people are working together in groups, such as the direction of gaze, posture (e.g., leaning toward or away from the activity area), and usage of tools (including pointing at or to the tool, as well as directly manipulating it). NICE was annotated by an author of the framework over Groups 1–3 and Group 5. This data is presented in .xlsx format.



**Figure 2** Multichannel (GAMR, NICE, speech transcription, and CPS) annotation “score” using ELAN (Brugman & Russel, 2004).

We extracted joint positions and orientations from each frame of the raw RGBD data, for all 32 joints on each body detected by Microsoft's body tracking SDK. This information (JSON) can be used to analyze body pose and gesture correlation to other modalities, or alone to classify gestures.

## 2.2 QUALITY CONTROL

By convention, participants are identified numerically from left (P1) to right (P3). Camera and microphone positioning are kept constant and the cameras calibrated using the standard Kinect SDK calibration procedure at the start of each session.

The raw data was recorded in .mkv format, including the depth channel, which is too large to include in the distributable dataset. We converted the RGB video to .mp4 and extracted the skeleton data from Azure depth channel.

In addition to individual .csv files, annotations have been loaded into .eaf files, which can be loaded in the ELAN environment (Brugman & Russel, 2004).

## 3 DATASET DESCRIPTION

**Object name** Weights Task Dataset

**Format names and versions** MP4, CSV, EAF, Excel, JSON

**Creation dates** 2022-09-22 — 2022-10-26

**Dataset creators** Ibrahim Khebour,<sup>1</sup> Richard Brutti,<sup>2</sup> Indrani Dey,<sup>3</sup> Rachel Dickler,<sup>4</sup> Kelsey Sikes,<sup>1</sup> Kenneth Lai,<sup>2</sup> Mariah Bradford,<sup>1</sup> Brittany Cates,<sup>1</sup> Paige Hansen,<sup>1</sup> Changsoo Jung,<sup>1</sup> Brett Wisniewski,<sup>1</sup> Corbyn Terpstra,<sup>1</sup> Leanne Hirshfield,<sup>4</sup> Sadhana Puntambekar,<sup>3</sup> Nathaniel Blanchard,<sup>1</sup> James Pustejovsky,<sup>2</sup> Nikhil Krishnaswamy<sup>1</sup>

**Language** English

**License** CC 4.0

**Repository name** Zenodo

**Publication date** 2023-09-27

## 4 REUSE POTENTIAL

This data was originally gathered to study multimodal indicators of CPS, but its rich multichannel nature also lends itself well to other lines of research. Researchers in education and learning sciences can use it to develop activities to support collaborative interaction and learning. Researchers in linguistics and psychology can use it to study interactive behavior and communication, including modeling the evolution of group common ground over time, a la Clark and Carlson (1981), and for natural language processing tasks such as assessing speech recognition fidelity (e.g., Terpstra et al., 2023, which compared the effects of different segmentation methods). The rich multimodality will be of use to researchers in AI. For example, the Kinect data can be used to develop and train gesture recognition algorithms (e.g., VanderHoeven, Blanchard, & Krishnaswamy, 2023) or object and action detectors. The different modalities can serve as signals to an interactive AI agent that assists facilitators and scale up collaborative group activities by interpreting key multimodal aspects of collaborative group interaction in context (cf. Bradford, Khebour, Blanchard, & Krishnaswamy, 2023). The dataset will continue to be updated at the public repository as additional annotations are performed, including of object positions, actions taken with the different objects, and of the common ground constructed between participants as the task unfolds.

<sup>1</sup> Colorado State University.

<sup>2</sup> Brandeis University.

<sup>3</sup> University of Wisconsin — Madison.

<sup>4</sup> University of Colorado Boulder.

Potential limitations or issues with reuse may include: while using the Azure (skeleton) data, the body IDs in some frames do not align with participant IDs, as the Microsoft tracker assigns a new body ID if it loses and regains a participant. The prosodic features, although useful in a number of applications, could introduce noise if during a single segmented utterance, more than one voice is actually talking at the same time.

Updates will be noted at the dataset link. The data is freely available for research purposes, as indicated in the consent form (also available at the dataset link).

## ACKNOWLEDGEMENTS

We gratefully acknowledge the participants in our study, who shall remain anonymous. Our thanks also go out to Dr. J. Ross Beveridge, who designed the original version of the Weights Task, which was modified to include the components in this dataset.

## FUNDING INFORMATION

This work was supported in part by the United States National Science Foundation (NSF) under grant #DRL 2019805 to the University of Colorado, with subcontracts to Colorado State University, Brandeis University, and the University of Wisconsin – Madison. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government. All errors and mistakes are, of course, the responsibilities of the authors.

## COMPETING INTERESTS

The authors have no competing interests to declare.

## AUTHOR CONTRIBUTIONS

**Ibrahim Khebour** — Data curation, Formal Analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing.

**Richard Brutti** — Methodology, Writing – original draft.

**Indrani Dey** — Data curation, Methodology, Writing – original draft.

**Rachel Dickler** — Conceptualization, Investigation, Methodology, Writing – original draft.

**Kelsey Sikes** — Data curation, Writing – original draft.

**Kenneth Lai** — Data curation, Methodology, Validation.

**Mariah Bradford** — Data curation, Methodology, Formal Analysis, Investigation, Validation, Writing – original draft.

**Brittany Cates** — Data curation, Writing – original draft.

**Paige Hansen** — Data curation, Writing – original draft.

**Changsoo Jung** — Data curation, Software.

**Brett Wisniewski** — Data curation.

**Corbyn Terpstra** — Data curation.

**Leanne Hirshfield** — Conceptualization, Funding acquisition, Supervision, Writing – original draft.

**Sadhana Puntambekar** — Conceptualization, Funding acquisition, Supervision, Writing – original draft.

**Nathaniel Blanchard** — Conceptualization, Supervision, Writing – original draft.

**James Pustejovsky** — Conceptualization, Funding acquisition, Supervision, Writing – original draft.

**Nikhil Krishnaswamy** — Conceptualization, Data curation, Supervision, Writing – original draft, Writing – review & editing.

## AUTHOR AFFILIATIONS

**Ibrahim Khebour**  [orcid.org/0009-0009-4374-7263](https://orcid.org/0009-0009-4374-7263)  
 Department of Computer Science, Colorado State University, Fort Collins, CO, USA

**Richard Brutti**  [orcid.org/0000-0003-0449-4418](https://orcid.org/0000-0003-0449-4418)  
 Department of Computer Science, Brandeis University, Waltham, MA, USA

**Indrani Dey**  [orcid.org/0009-0000-9284-6078](https://orcid.org/0009-0000-9284-6078)  
 Department of Educational Psychology, University of Wisconsin – Madison, Madison, WI, USA

**Rachel Dickler**  [orcid.org/0000-0002-9018-4848](https://orcid.org/0000-0002-9018-4848)  
 Institute of Cognitive Science, University of Colorado, Boulder, CO, USA

**Kelsey Sikes**  [orcid.org/0009-0003-9711-920X](https://orcid.org/0009-0003-9711-920X)  
 Department of Computer Science, Colorado State University, Fort Collins, CO, USA

**Kenneth Lai**  [orcid.org/0000-0003-2870-7019](https://orcid.org/0000-0003-2870-7019)  
 Department of Computer Science, Brandeis University, Waltham, MA, USA

**Mariah Bradford**  [orcid.org/0009-0009-2162-3307](https://orcid.org/0009-0009-2162-3307)  
 Department of Computer Science, Colorado State University, Fort Collins, CO, USA

**Brittany Cates**  [orcid.org/0009-0000-4169-0616](https://orcid.org/0009-0000-4169-0616)  
 Department of Computer Science, Colorado State University, Fort Collins, CO, USA

**Paige Hansen**  [orcid.org/0009-0009-7350-2312](https://orcid.org/0009-0009-7350-2312)  
 Department of Computer Science, Colorado State University, Fort Collins, CO, USA

**Changsoo Jung**  [orcid.org/0000-0002-2232-4300](https://orcid.org/0000-0002-2232-4300)  
 Department of Computer Science, Colorado State University, Fort Collins, CO, USA

**Brett Wisniewski**  [orcid.org/0009-0005-1236-069X](https://orcid.org/0009-0005-1236-069X)  
 Department of Computer Science, Colorado State University, Fort Collins, CO, USA

**Corbyn Terpstra**  [orcid.org/0009-0006-7005-8437](https://orcid.org/0009-0006-7005-8437)  
 Department of Computer Science, Colorado State University, Fort Collins, CO, USA

**Leanne Hirshfield**  [orcid.org/0000-0003-0111-6948](https://orcid.org/0000-0003-0111-6948)  
 Institute of Cognitive Science, University of Colorado, Boulder, CO, USA

**Sadhana Puntambekar**  [orcid.org/0000-0002-7102-0127](https://orcid.org/0000-0002-7102-0127)  
 Department of Educational Psychology, University of Wisconsin – Madison, Madison, WI, USA

**Nathaniel Blanchard**  [orcid.org/0000-0002-2653-0873](https://orcid.org/0000-0002-2653-0873)  
 Department of Computer Science, Colorado State University, Fort Collins, CO, USA

**James Pustejovsky**  [orcid.org/0000-0003-2233-9761](https://orcid.org/0000-0003-2233-9761)  
 Department of Computer Science, Brandeis University, Waltham, MA, USA

**Nikhil Krishnaswamy**  [orcid.org/0000-0001-7878-7227](https://orcid.org/0000-0001-7878-7227)  
 Department of Computer Science, Colorado State University, Fort Collins, CO, USA

Khebour et al.  
*Journal of Open  
 Humanities Data*  
 DOI: 10.5334/johd.168

## REFERENCES

**Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... Weinert, R.** (1991). The hcrc map task corpus. *Language and Speech*, 34(4), 351–366. DOI: <https://doi.org/10.1177/002383099103400404>

**Bradford, M., Khebour, I., Blanchard, N., & Krishnaswamy, N.** (2023). Automatic detection of collaborative states in small groups using multimodal features. In *Proceedings of the 24th international conference on artificial intelligence in education*. DOI: [https://doi.org/10.1007/978-3-031-36272-9\\_69](https://doi.org/10.1007/978-3-031-36272-9_69)

**Brugman, H., & Russel, A.** (2004, May). Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the fourth international conference on language resources and evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA). Retrieved from <http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf>

**Brutti, R., Donatelli, L., Lai, K., & Pustejovsky, J.** (2022, June). Abstract Meaning Representation for gesture. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 1576–1583). Marseille, France: European Language Resources Association. Retrieved from <https://aclanthology.org/2022.lrec-1.169>

**Clark, H. H., & Carlson, T. B.** (1981). Context for comprehension. *Attention and performance IX*, 313, 30.

**Dey, I., Puntambekar, S., Li, R., Gengler, D., Dickler, R., Hirshfield, L. M., ... Krishnaswamy, N.** (2023). The NICE framework: analyzing students' nonverbal interactions during collaborative learning. In *Pre-conference workshop on collaboration analytics at 13th international learning analytics and knowledge conference (lak 2023)*. DOI: <https://doi.org/10.22318/cscl2023.218179>

**Gillies, R. M.** (2008). The effects of cooperative learning on junior high school students' behaviours, discourse and learning during a science-based learning activity. *School Psychology International*, 29(3), 328–347. DOI: <https://doi.org/10.1177/0143034308093673>

**Langer-Osuna, J. M., Gargroetzi, E., Munson, J., & Chavez, R.** (2020). Exploring the role of off-task activity on students' collaborative dynamics. *Journal of Educational Psychology*, 112(3), 514. DOI: <https://doi.org/10.1037/edu0000464>

**Liu, B., Cai, H., Ji, X., & Liu, H.** (2017). Human-human interaction recognition based on spatial and motion trend feature. In *2017 ieee international conference on image processing (icip)* (pp. 4547–4551). DOI: <https://doi.org/10.1109/ICIP.2017.8297143>

**Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I.** (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

**Roschelle, J., & Teasley, S. D.** (1995). The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning* (pp. 69–97). DOI: [https://doi.org/10.1007/978-3-642-85098-1\\_5](https://doi.org/10.1007/978-3-642-85098-1_5)

**Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S.** (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143, 103672. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0360131519302258>. DOI: <https://doi.org/10.1016/j.compedu.2019.103672>

**Terpstra, C., Khebour, I., Bradford, M., Wisniewski, B., Krishnaswamy, N., & Blanchard, N.** (2023). *How good is automatic segmentation as a multimodal discourse annotation aid?*

**Van Gemeren, C., Poppe, R., & Veltkamp, R. C.** (2016). Spatio-temporal detection of fine-grained dyadic human interactions. In *Human behavior understanding: 7th international workshop, hbu 2016, Amsterdam, the Netherlands, October 16, 2016, proceedings* 7 (pp. 116–133). DOI: [https://doi.org/10.1007/978-3-319-46843-3\\_8](https://doi.org/10.1007/978-3-319-46843-3_8)

**VanderHoeven, H., Blanchard, N., & Krishnaswamy, N.** (2023). Robust motion recognition using gesture phase annotation. In *International conference on human-computer interaction* (pp. 592–608). DOI: [https://doi.org/10.1007/978-3-031-35741-1\\_42](https://doi.org/10.1007/978-3-031-35741-1_42)

**Velikovich, L., Williams, I., Scheiner, J., Aleksic, P. S., Moreno, P. J., & Riley, M.** (2018). Semantic lattice processing in contextual automatic speech recognition for google assistant. In *Interspeech* (pp. 2222–2226). DOI: <https://doi.org/10.21437/Interspeech.2018-2453>

**Wang, I., Fraj, M. B., Narayana, P., Patil, D., Mulay, G., Bangar, R., ... Ruiz, J.** (2017). Eggnog: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *2017 12th Ieee international conference on automatic face & gesture recognition (fg 2017)* (pp. 414–421). DOI: <https://doi.org/10.1109/FG.2017.7145>

**Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., & Samaras, D.** (2012). Two-person interaction detection using body-pose features and multiple instance learning. In *2012 Ieee computer society conference on computer vision and pattern recognition workshops* (pp. 28–35). DOI: <https://doi.org/10.1109/CVPRW.2012.6239234>

Khebour et al.  
*Journal of Open  
Humanities Data*  
DOI: 10.5334/johd.168

7

#### TO CITE THIS ARTICLE:

Khebour, I., Brutti, R., Dey, I., Dickler, R., Sikes, K., Lai, K., Bradford, M., Cates, B., Hansen, P., Jung, C., Wisniewski, B., Terpstra, C., Hirshfield, L., Puntambekar, S., Blanchard, N., Pustejovsky, J., & Krishnaswamy, N. (2024). When Text and Speech are Not Enough: A Multimodal Dataset of Collaboration in a Situated Task. *Journal of Open Humanities Data*, 10, 7, pp. 1–7. DOI: <https://doi.org/10.5334/johd.168>

**Submitted:** 14 October 2023

**Accepted:** 05 December 2023

**Published:** 17 January 2024

#### COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

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