

Datasheet for Assemblage

CHANG LIU*, Syracuse University

REBECCA SAUL*, Booz Allen Hamilton

YIHAO SUN, Syracuse University

EDWARD RAFF, Booz Allen Hamilton; University of Maryland, Baltimore County

MAYA FUCHS, Booz Allen Hamilton

TOWNSEND SOUTHARD PANTANO, Syracuse University

JAMES HOLT, Laboratory for Physical Sciences

KRISTOPHER MICINSKI, Syracuse University

1 MOTIVATION

- **For what purpose was the dataset created?**

Assemblage was built to provide an open-source collection of ground-truth for binary analysis. It is meant to provide a common baseline for benchmarking results in binary analysis, particularly apropos Windows PE and Linux ELF binaries.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

Syracuse University created the dataset through a partnership with Booz Allen Hamilton.

2 COMPOSITION

- **What do the instances that comprise the dataset represent?**

The dataset comprises source code crawled from GitHub and vcpkg. For each source repository, the dataset includes a collection of sibling binaries, each built using a different toolchain. The dataset includes source-to-binary mappings, to anticipate its usage in tasks which necessitate a materialized relationship between sources and binaries at a basic-block-level granularity.

- **How many instances are there in total (of each type, if appropriate)?**

We distribute a “recipe” to build a dataset of 1,536,171 binaries (Windows PE and Linux); our publicly-released dataset (permissive licenses) contains 382,955 binaries.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).**

Our tool attempts to build Windows PE and Linux ELF binaries from crawling GitHub. This process may fail due to both (a) circumstances outside of our control (software which does not build) and (b) processes we have yet to support (e.g., requiring a system library or setup which our tools do not anticipate). Please see our paper for a detailed characterization of build limitations. The data is complete in the sense that, whenever there is a metadata entry, there is a corresponding artifact.

- **What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.**

The dataset contains 32 and 64-bit Windows PE and Linux ELF executables, and corresponding metadata. This metadata, in the form of an SQLite database, includes build provenance for each binary along with a detailed source-to-binary mapping in the form of both (a) references from the source into the binary and (b) Program DataBase (PDB) files. In our experience, these PDB files facilitate significantly better performance of decompilers (such as IDA Pro and Ghidra).

- **Is there a label or target associated with each instance?**

Yes, our SQLite database relates each binary artifact with its provenance.

- **Is any information missing from individual instances?**

*Equal contributions.

Authors’ addresses: Chang Liu*, Syracuse University; Rebecca Saul*, Booz Allen Hamilton; Yihao Sun, Syracuse University; Edward Raff, Booz Allen Hamilton; University of Maryland, Baltimore County; Maya Fuchs, Booz Allen Hamilton; Townsend Southard Pantano, Syracuse University; James Holt, Laboratory for Physical Sciences; Kristopher Micinski, Syracuse University.

Not to our knowledge, this would constitute a bug on our part.

- **Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?**

Yes, our SQLite database relates binaries, sources, and build / configuration information. However, we currently only examine repositories in isolation, and recover only a single version for each repository; we look to support multiple versions of each repository in future work. Similarly, we omit fine-grained author / attribution.

- **Are there recommended data splits (e.g., training, development/validation, testing)?**

No, we do not have recommended training / test splits. Our experience is that the Assemblage dataset represents a superset of available binaries; applying the dataset to any specific task will likely require additional data cleaning, down-selection, and similar tasks.

- **Are there any errors, sources of noise, or redundancies in the dataset?**

We do not believe that there are any mis-labelings of source code. It is possible that binaries will not build, indeed, this is common—our paper and its appendix offer a characterization of failure rates and build cost. Sources of noise include the expected noise of code on GitHub, which is not necessarily balanced and cannot be proven with certainty to be correct, bug-free, or even benign. We believe our dataset does not contain redundancies on a per-binary level (only one binary per build configuration) and construe opposition to this as a bug.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer**

The binary and source code in our dataset is self-contained. We also distribute “recipes” which facilitate reproducing a corpus from sources. This process has potential for failure due to changes over time (repositories becoming private or deleted).

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)?**

No, except for the fact that we reproduce publicly-available sources on GitHub, whose terms of service forbid such material.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

No, other than the extent to which software source code may contain offensive language or content.

3 COLLECTION PROCESS

- **How was the data associated with each instance acquired?**

We crawled a total of 4 million C and C++ repositories from GitHub, whose repositories originated in years ranging from 2008 to 2023.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?**

We use Make, GNU, Clang, Visual Studio and other compiler toolchains and related software, either proprietary or distributed as free software, such as MSBuild, CMake and Make. We also used AWS’s command line interface, compressing software such as 7-zip and Zip. On APIs, we used GitHub’s API and vcpkg’s command line interface.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

This dataset is not a sample from a larger dataset, but the source it built from is a subset of the set of all C++ binaries living on GitHub, other repositories are the ones either failed to build, or not visible to user, or GitHub’s API didn’t index them.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Graduate employees paid throughout the course of the project were paid via graduate stipends at Syracuse University, and also via hourly pay. Undergraduates were paid summer wage through an Research Experiences for Undergraduates (REU) program.

- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.**

Our final dataset was created during 2023. We ran ASSEMBLAGE from May 2022 through December 2023, with the bulk of our reported dataset being collected during a span of 20 weeks in 2023. Repositories were crawled as they were built to ensure we got a copy that was as recent as possible.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?**

An IRB was not involved as this research involved no human subjects or similarly-interested parties. We have worked to ensure high ethical standards in building Assemblage to respect rate limits and avoid the distribution of unlicensed code (while allowing corpus summaries to be distributed in the form of “recipes”).

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

We collected the data through GitHub, through its API.

- **Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.**

No, we did not inform the authors we built and distributed binaries corresponding to their source artifacts. We worked to ensure we only distribute binaries which have permissive licenses as determined via GitHub’s API.

- **Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.**

We used only open-sourced libraries where the authors have pre-consented publicly to the use of their code via their selected license.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).**

This question is not applicable to our work as the data used has all been pre-consented for public use/consumption via OSS licenses. We are careful to make sure that our usage and distribution of recipes comply with those licenses.

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.**

Our work has no impact on any human subjects and is using already existing and openly available artifacts. Thus, no further analysis was deemed applicable.

- **Any other comments?**

No.

4 PREPROCESSING/CLEANING/LABELING

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.**

Before attempting the building of binaries, we parse and modify certain compiler flags stored in either Makefile or MSVC related config files, to produce diver binary from one copy of source code.

- **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.**

Yes, the unmodified binaries and pdb files (if applicable) are distributed in our dataset, which can be accessed the link <https://assemblagedocs.readthedocs.io/en/latest/dataset.html#dataset-access>.

- **Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.**

The tool used to build the dataset is available: <https://github.com/Assemblage-Dataset/Assemblage>. We used Dia2dump to process the binary with PDB files, the link for Dia2dump is here <https://learn.microsoft.com/en-us/visualstudio/debugger/debug-interface-access/dia2dump-sample?view=vs-2022>.

- **Any other comments?**

No.

5 USES

- **Has the dataset been used for any tasks already? If so, please provide a description.**

Yes, this dataset has been used for work in binary similarity, decompilation, compiler provenance and similar binary analysis tasks. Our full paper gives more details as to the dataset’s application to modern binary analysis tasks.

- **Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.**

We have a website with links to our code and papers: assemblage-dataset.net. We also have a detailed page about the dataset that constantly updates, <https://assemblagedocs.readthedocs.io/en/latest/dataset.html#dataset-access>.

- **What (other) tasks could the dataset be used for?**

We anticipate the broad potential for application to tasks in reasoning over (corpora of) binaries, including problems in decompilation, malware analysis, security auditing, and similar tasks. Especially with the rise of large language models, we hope the dataset—with its rich source-to-binary mapping—will facilitate new binary analysis results using novel architectures.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?**

Not to our knowledge.

- **Are there tasks for which the dataset should not be used? If so, please provide a description.**

Not to our knowledge.

- **Any other comments?**

No.

6 DISTRIBUTION

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.**

The public portion of dataset is released and accessible by everyone, the links can be found at the link <https://assemblagedocs.readthedocs.io/en/latest/dataset.html#dataset-access>.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?**

The dataset is distributed as compressed files, hosted on Hugging Face and Kaggle. The source code is available on GitHub. We do not have a DOI at this time.

- **When will the dataset be distributed?**

The public portions of the dataset are publicly available now on Kaggle and Hugging Face.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.**

ASSEMBLAGE’s code is available under the MIT licenses. Individual binaries and source repositories have their own licenses, which our metadata records.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.**

No, the dataset is public and has no IP restrictions other than the licenses pertaining to the artifacts.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.**

No, beyond the export controls that would pertain to the source artifacts in our corpus, which is drawn from GitHub.

- **Any other comments?**

Our source code and replication recipe can be accessed at <https://github.com/Assemblage-Dataset/Assemblage>.

7 MAINTENANCE

- **Who will be supporting/hosting/maintaining the dataset?**

Kristopher Micinski's research team at Syracuse University will be supporting the long-term development and maintenance.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

Lead PI is Kristopher Micinski; email may be sent to kkmicins@syr.edu

- **Is there an erratum? If so, please provide a link or other access point.**

We will post errata on the main GitHub page for our Assemblage repository: <https://github.com/Assemblage-Dataset/Assemblage>

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?**

We plan to keep working on Assemblage indefinitely, expanding its capabilities. At this time, we have not committed to a release schedule; we anticipate releasing updates to the dataset a few times a year.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.**

The dataset does not relate to people, we currently do not provide author-level metadata.

- **Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.**

We will keep the current version of our dataset public and hosted in a publicly-available manner.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.**

We welcome collaborators to extend both the dataset itself and the ASSEMBLAGE tool used to create it. We will validate changes to ensure integrity of the dataset and freedom from malicious or otherwise-offensive material. We do not currently have a mechanism to communicate dataset changes to dataset consumers, we are looking into platforms such as community mailing lists.

- **Any other comments?**

No.