# Enabling useful data sharing through format identification and text mining

Dimitrios-Georgios Akestoridis
University of Ioannina
akestoridis@gmail.com

Tristan Henderson
University of St Andrews
tnhh@st-andrews.ac.uk

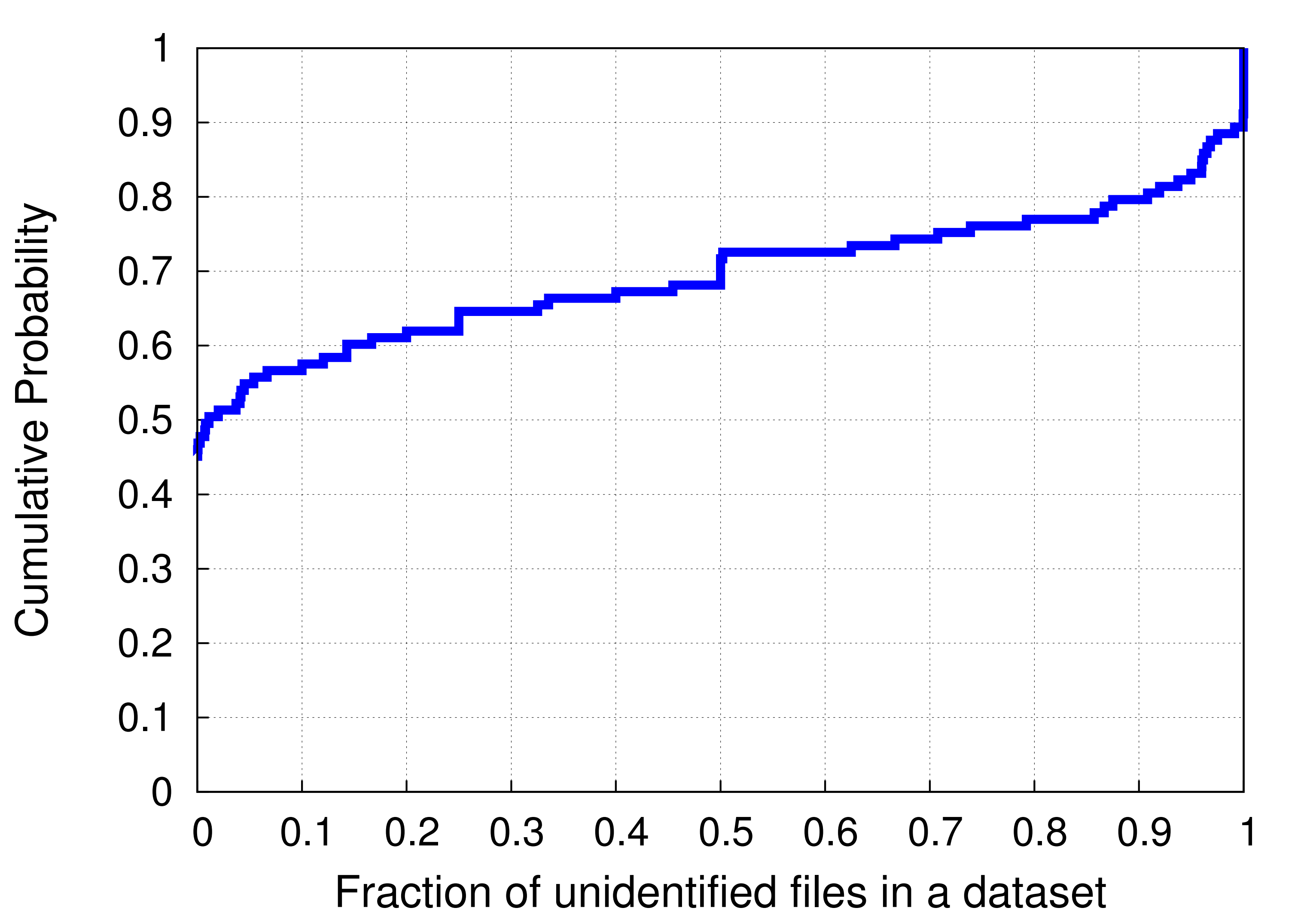
## 1. Introduction

We have run CRAWDAD (Community Resource for Archiving Wireless Data At Dartmouth) since 2005 [7], and it is now the leading wireless network data archive, with over 130 datasets and tools used for over 2,100 papers by 9,000 users from 116 countries. As one of the most obvious benefits of sharing data is that data are used by other researchers, we have observed that our datasets are used by researchers from many other fields besides computer science, including geography, sociology, biology and zoology. Researchers from other fields may have different requirements. In this work we look at how data collectors can share data such that they can be easily understood and used by researchers, both in the same and other fields. In particular we look at which file formats researchers use, by comparing the file formats used by data contributors to those utilised by data users, and describe our new workflow for creating new derived datasets to maximise the utility of the datasets in our data archive.

## 2. File formats used by data collectors

To see what file formats are used by data collectors, we ran the file format identification tool DROID [2] on all of the files in the CRAWDAD data archive. DROID was unable to identify the majority (57.15%) of the files. Of the files that were identified, 94.96% were identified by extension, which is the least reliable identification method. This might indicate that the data collectors in our archive have used their own custom formats.

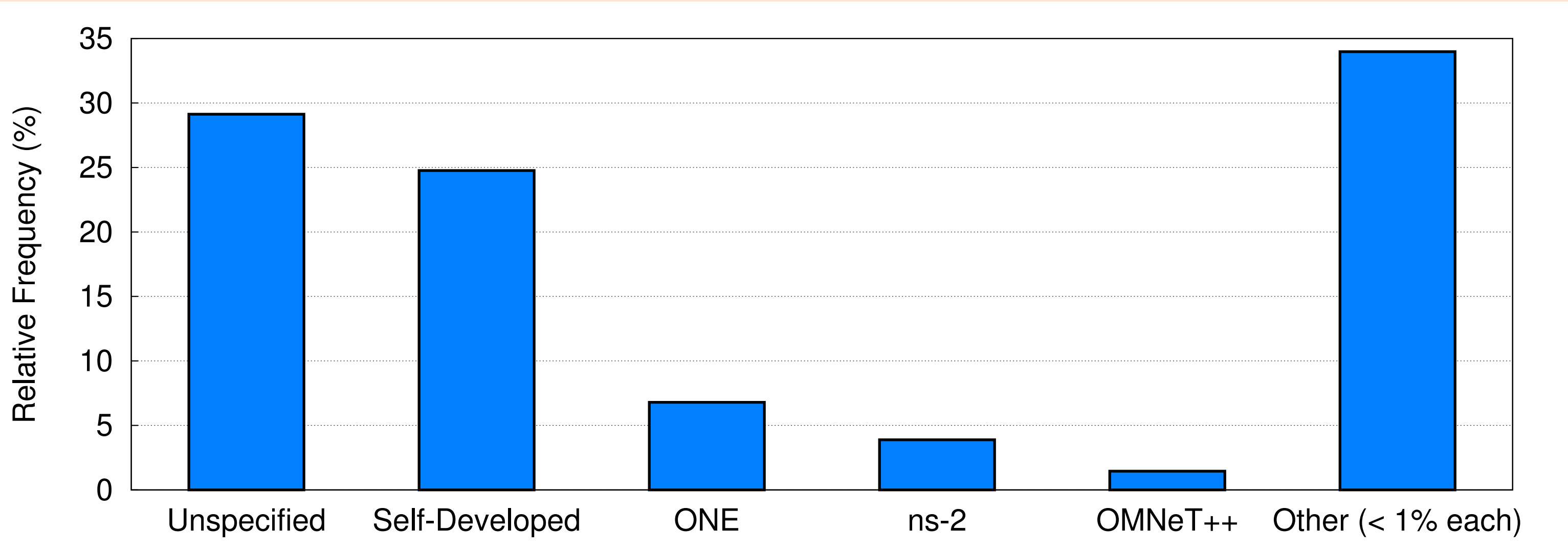
Although many individual files were unable to be identified, when we aggregate the files on a per-dataset basis, DROID was able to identify at least half of the files in 71.6% of the datasets.



Plain text, CSV (comma-separated value), and pcap (network packet capture) files were the most commonly identified files among the datasets. Files without any extension were by far the most commonly unidentified files among the datasets.

## 3. File formats used by data users

To see which file formats are preferred by data users, we used GATE [4] to extract information from a corpus of papers that use CRAWDAD datasets [5] to discover which software tools are used in these research papers. GATE uses a pattern-matching engine called JAPE to identify named entities in a document, and we created a JAPE grammar to capture statements that refer to tools. This grammar extracted the relevant parts of the papers, and reduced the number of words that we needed to manually inspect by 83.6%. The following bar chart illustrates the relative frequency of software tools that we identified in 150 papers from the corpus.



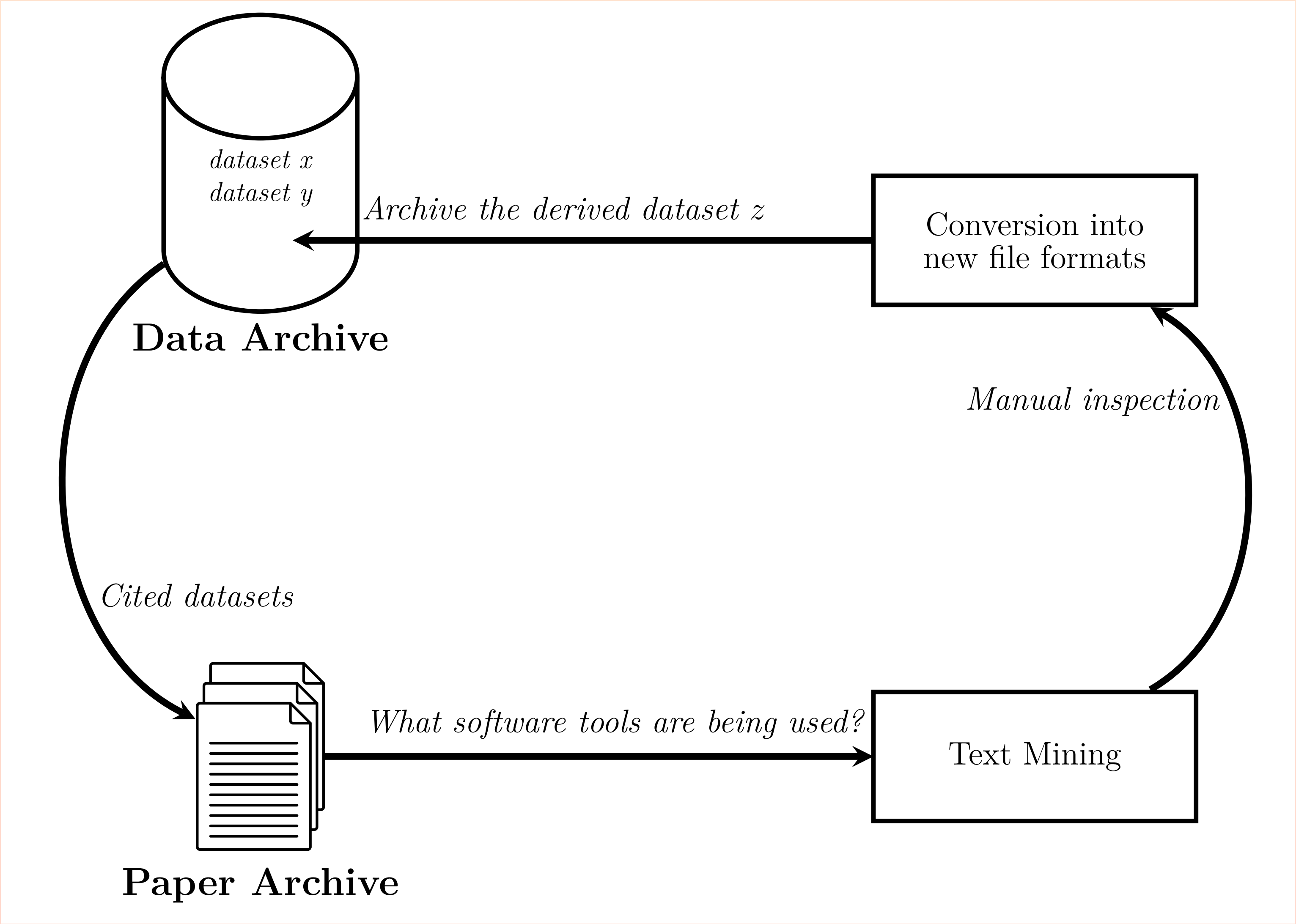
Unfortunately, 60 research papers did not provide sufficient information for us to identify the tools used in their analyses. The authors of 47 research papers developed their own custom tools. Of the remaining papers where we could identify tools, we found 62 different tools in use. Common tools included the ONE [6], ns-2 [8] and OMNeT++ [9] network simulators. These all use different file formats to the data provided by the original data collectors, and converting files to these formats can involve additional processing and assumptions. For instance, 802.11 wireless measurements in the CRAWDAD datasets can be converted to opportunistic network contact files for ONE, but this needs assumptions to be made about mobility and communication range.

Our analysis leads to two recommendations for improving the reproducibility and repeatability of experiments:

1. standard descriptors for tools might be useful
2. data archives could provide data in formats used by other researchers as well as the original data collectors.

## 4. Archiving of derived datasets

If data archives are to provide new formats for collected data, then it is important that these datasets are documented and persistently identified to enable data citation in line with the FORCE11 principles [3]. We propose to consider these new formats as derived datasets. We already provide persistent identifiers (DOIs) for each CRAWDAD dataset, but DataCite suggest new DOIs for derived datasets and provide definitions for relations between original and derived datasets [1]. We are currently developing a new workflow for analysing new research papers to determine which formats need to be created, adding these derived datasets to the archive and providing appropriate citation and relationship information.



*dataset x*
*dataset y*
Data Archive
*Archive the derived dataset z*
Conversion into new file formats
*Manual inspection*
*Cited datasets*
Paper Archive
*What software tools are being used?*
Text Mining

## 5. Acknowledgements

We are grateful for the support of our current funders ACM SIGMOBILE, and our past funders ACM SIGCOMM, Aruba Networks, Intel and the National Science Foundation.

## References

[1] DataCite Metadata Working Group. DataCite metadata schema for the publication and citation of research data, 2015. doi:10.5438/0010.
[2] Digital record object identification tool (DROID). Online at http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/.
[3] FORCE11 Data Citation Synthesis Group. Joint declaration of data citation principles. Online at https://www.force11.org/group/joint-declaration-data-citation-principles-final.
[4] GATE: General architecture for text engineering. Online at https://gate.ac.uk/.
[5] T. Henderson and D. Kotz. CRAWDAD wireless network data citation bibliography, 02 Jan. 2015. doi:10.6084/m9.figshare.1203646.
[6] A. Keränen, J. Ott, and T. Kärkkäinen. The ONE Simulator for DTN Protocol Evaluation. In *SIMUTools '09: Proceedings of the 2nd International Conference on Simulation Tools and Techniques*, Rome, Italy, 2009. doi:10.4108/ICST.SIMUTOOLS2009.5674.
[7] D. Kotz and T. Henderson. CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth. *IEEE Pervasive Computing*, 4(4):12–14, Oct. 2005. doi:10.1109/mprv.2005.75.
[8] The network simulator - ns-2. Online at http://www.isi.edu/nsnam/ns/.
[9] OMNeT++ discrete event simulator. Online at https://omnetpp.org/.

*Repository Fringe 2016, Edinburgh, UK, August 2016*