

# FUSE: A Reproducible, Internet-scale Dataset of Spreadsheets

Titus Barik<sup>\*†</sup>, Kevin Lubick<sup>†</sup>, Justin Smith<sup>†</sup>, John Slankas<sup>†</sup>, Emerson Murphy-Hill<sup>†</sup>

<sup>\*</sup>ABB Corporate Research, Raleigh, North Carolina, USA

<sup>†</sup>North Carolina State University, Raleigh, North Carolina USA

titus.barik@us.abb.com, {kjlubick, jssmit11, jbslanka}@ncsu.edu, emerson@csc.ncsu.edu

**Abstract**—We submit a corpus consisting of 1.5 million spreadsheets, extracted using the Common Crawl corpus (based on Blekko). This corpus is compared against a proprietary index from a leading search engine company to measure representative against other commercial index. This corpus is intended to replace EUSES.

## I. INTRODUCTION

End-user programmers today constitute a broad class of users, including teachers, administrators research scientists, and even children. Although these users are typically not professional software developers, their roles routinely involve computational tasks, that, in many ways, are similar to those of developers — even at a cognitive level. Perhaps the most common form of end-user programming software are spreadsheets.

[1] [2] [3] [4] [5]

Spreadsheets are one of the most common end user programming environments [6]. Spreadsheets provide a platform for formula-based and conditional computations. Within a spreadsheet, users can develop complex programs using macros. Like traditional programming environments, researchers have studied errors, code smells, refactoring, and debugging in spreadsheets [7], [8], [2], [9].

Researchers have curated several spreadsheet corpora to support such studies. [10], [9], [11]

Such research is supported by some notable existing corpora, such as the EUSES spreadsheet corpus [12] and the Enron spreadsheet dataset [13]. However, the available spreadsheet corpora exhibit several limitations. Currently spreadsheet corpora exist as snapshots; they do not provide mechanisms for reproduction or updates. Further, with 4498 spreadsheets, the EUSES dataset represents a small sample of all existing spreadsheets and the Enron dataset contains spreadsheets from only one company. Researchers who wish to use these corpora must manually reason about the properties of each spreadsheet as neither corpus provides spreadsheet-level metadata. For example, a researcher only interested in studying spreadsheet macros would have to download and inspect the contents of each spreadsheet from the EUSES corpus even though only 126 contain macros.

Alternatively, we provide a corpus creation technique that overcomes many of these limitations. Not only is our corpus larger, but our technique is open and reproducible. We publicly release the scanning software used to create this corpus.

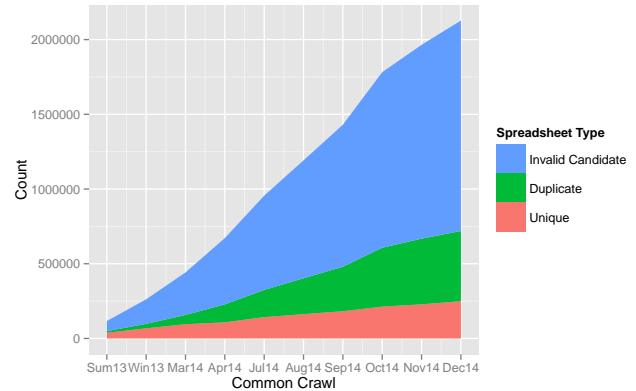


Fig. 1. Cumulative count of candidate spreadsheets with each additional crawl. One problem is that unless the diversity of the crawl increases, unique spreadsheets are reaching a local maximum.

Accordingly, we encourage researchers to validate our data and methods. Further, to help researchers navigate the dataset, we provide metadata for each spreadsheet in our corpus as well as the tool used to compute that metadata.

Description of data should go here.

The contributions of this data set are:

- 1) A
- 2) A corpus of ??? spreadsheets pulled from the pubic web
- 3) A pipeline of tools that allows other researchers to more easily perform spreadsheet extraction and analysis at scale
- 4) A detailed set of metadata for this corpus and two other corpora [foo], [bar] which allow researchers to make queries without having to download or analyze the entire spreadsheet corpus.

similar in size to other large corpora representativeness against Google index metadata

helps with tool evaluation

accessible – getting this for yourself is a high cost maybe 5k

this set is reproducible compared with existing data sets, it is new – the others are more than a decade old.

even the enron corpus is 15000 spreadsheets – whoopee. But it's good to compare against since it's a private corpus. what metrics?

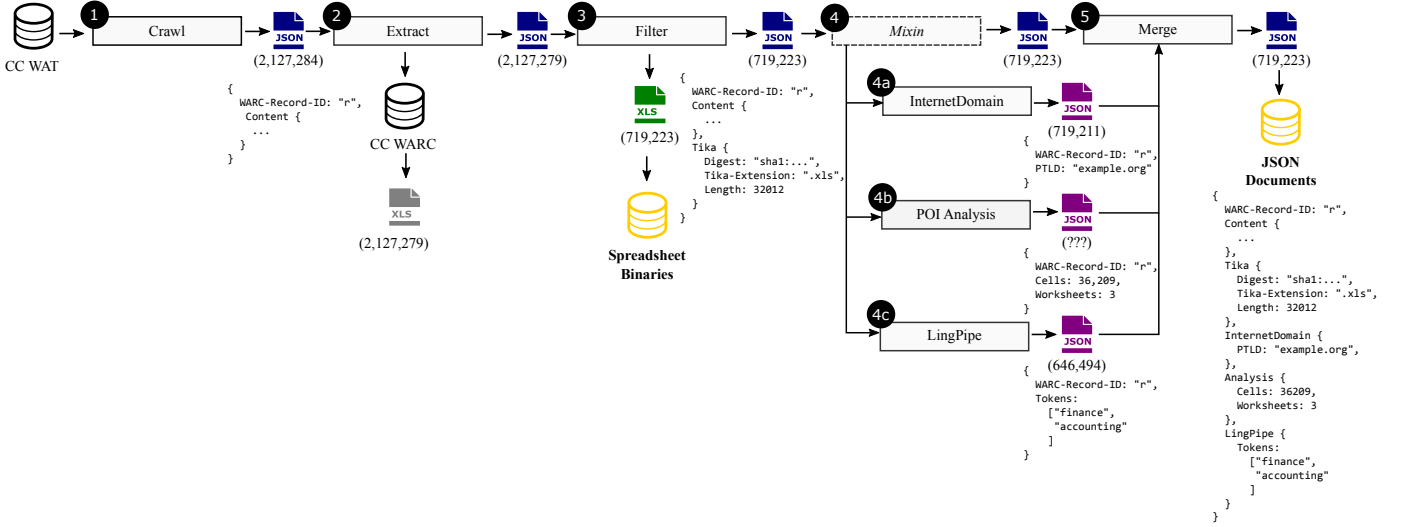


Fig. 2. The MapReduce pipeline for extracting spreadsheets and associated spreadsheet metadata from Common Crawl. In CRAWL (Stage 1), WAT segments containing HTTP headers and offset information into WARC records are parsed. Records that heuristically match spreadsheets (e.g., a Content-Type of application/vnd.ms-excel) are retained. In EXTRACT (Stage 2), ...

TABLE I  
COMMON CRAWL ARCHIVE

Description	TB	Yield
Summer 2013	30.6	42.14%
Winter 2013	35.1	33.46%
March 2014	36.4	
April 2014	41.2	
July 2014	59.2	
August 2014	46.6	
September 2014	48.9	
October 2014	59.1	
November 2014	31.4	
December 2014	35.2	

## II. METHODOLOGY

We selected the Common Crawl<sup>1</sup> index as the primary source for our spreadsheet corpus, because it contains over ???TB of publicly available web crawl data and is regularly updated.

To extract spreadsheets from this index, first we crawled all the available WAT files in the index targeting files that could potentially be spreadsheets, including files tagged with MSDN content types, and files with extensions containing “.xls”. This crawl was the most computationally intensive task in our pipeline, consuming approximately ??? hours of CPU time. The crawl identified ??? candidate spreadsheets which we extracted from their associated WARC files.

Because the WAT files contained some incomplete and incorrect tags, we extracted some invalid spreadsheets in the first phase. We used Tika<sup>2</sup> to identify the valid spreadsheets and then filtered out the invalid files. This stage resulted in ??? valid spreadsheets.

Next we processed the valid spreadsheets using Apache POI<sup>3</sup> to extract metadata from each spreadsheet (see Section III). By computing a sha512 hash we identified and removed duplicate spreadsheets, resulting in a final total of ??? spreadsheets.

+ TODO(tbarik): Copy Mixin and JsonMerge, not shown, but utility tool to easily copy sets and merge them.

### A. MapReduce Pipeline

- 1) *Crawl*: In this step,
  - 2) *Extract*: In this step,
  - 3) *Filter*: In this step,
  - 4) *Mixin*: In this step,
  - 5) *Merge*: In this step, limitation: truncation, 2 MB, 5 MB?
- + TODO(tbarik): Stages or tasks? needed to create WAT index files

For Summer 2013, Winter 2013, and March 2013, no pre-created WAT path files were available. To manually create this we did:

## III. DATA SCHEMA

In this section, we describe the schema. The metadata we collected for the indices was largely influenced by the summary statistics presented in [12]. For each spreadsheet, there are over 450 entries, so we will not list them all here. In general, the entries summarize the contents of the cells. To list a few examples, the number of times a given Excel function (such as SUM or VLOOKUP) is used, the total number of input or data cells, the number of numeric input cells, the number of formulas used more than 50 times, the most common formula used, etc.

render json

<sup>1</sup><http://commoncrawl.org>

<sup>2</sup><http://tika.apache.org>

<sup>3</sup><http://poi.apache.org>

asdfsdf

#### A. Mixins

how do you get it?  
csv file mongodb recordobject warc extracts  
how did we decide to filter? MSDN content-type  
google obtained from proprietary database query that ex-  
tracted from google index all spreadsheets with content type  
application/ms-excel.  
common crawl segments -i subproblems  
docjoiner  
cleaning  
put lingpipe put interndomain put analysisstool

#### IV. DESCRIPTION OF SPREADSHEET CORPUS

```
db.s.aggregate(  
  { "$project":  
    { "InternetDomainName.Public-Suffix" : true },  
  { "$group":  
    { "_id": "$InternetDomainName.Public-Suffix",  
    { "$sort": { "count" : -1 } },  
    { "$limit": 10 }  
  })
```

since it's sdescription of corpus, should be descriptive?  
counts avg etc.

absfrequency, relativefrequency gov, 250211 org, 171845  
com, 112573 edu, 63572 gov.au, 43490 pa.us, 7621 net, 7409  
mn.us, 4641 ac.uk, 3423 ca.us, 3003

absfreq, relfreq census.gov, 143135 triathlon.org, 106486  
amamanualofstyle.com, 45118 abs.gov.au, 42941 utah.gov,  
22242 ohio.gov, 16739 usda.gov, 13016 worldbank.org, 11062  
theahl.com, 10350 eia.gov, 8216

#### RQ: How many domains are represented?

4,381 domains. After dedup, 4,342.

RQ what types of headers do people use?

how canonical are urls?

not very

results from this can be used to guide future crawls to  
increase the diversity of the spreadsheet corpus.

Actually looking at binaries, so this looks at dedup:

application/vnd.ms-excel, 238673  
application/vnd.openxmlformats-  
officedocument.spreadsheetml.sheet,  
10555 application/vnd.openxmlformats-  
officedocument.spreadsheetml.template, 148

all zero: application/vnd.ms-excel.sheet.macroEnabled.12  
application/vnd.ms-excel.template.macroEnabled.12  
application/vnd.ms-excel.addin.macroEnabled.12  
application/vnd.ms-excel.sheet.binary.macroEnabled.12

RQ: How much can you trust HTTP headers?

RQ: Train a text classifier to identify topicality.

RQ: Evolution of spreadsheets

Breakdown of analysis:

"OK", 220760 "BIFF5", 17643 "OTHER", 10782 "COR-  
RUPT", 129 "ENCRYPTED", 62

TABLE II  
SELECTED FUNCTIONS WITH COUNTS PER 1000 FORMULA CELLS

	FUSE	Enron	EUSES
IF	178.8	156.9	166.8
+ (operator)	166.4	217.5	167.6
SUM	87.7	80.3	153.6
ISBLANK	57.8	0.1	27.4
VLOOKUP	30.8	52.3	12.2
HLOOKUP	9.5	2.8	1.4
AVERAGE	32.0	9.2	7.9
AND	6.6	15.9	21.7
NOT	5.5	0.0	1.9

Total Input cells: 357210294

Total Formula cells: 10776903

Total Non-empty cells: 357210294 + 10776903 =  
367987197 Average non-empty cells per workbook:  
367987197/220760 = 1667

Number of workbooks with formulas: 14782

Number of formulas/workbook with formulas: 10776903/  
14782 = 729  
Count: { "\$sum": 1 } } },

Number of unique formulas: 894361

Number of unique formulas/workbook with formulas:  
894361/ 14782 = 60

Number of different functions used: 219

Total Sheets: 346247

Maximum number of sheets: 147

Programming in spreadsheets:

154596 of our unique formulas contained IF, or one of  
its cousins like SUMIF, COUNTIF. 150461 of our unique  
formulas contained IF (and maybe other functions). 44584 of  
our unique formulas contained IF two or more times. 9816  
unique formulas used IF 5 or more times, typically in a  
nested fashion 302 unique formulas used IF ten or more times,  
typically in a nested fashion. This may indicate a need for a  
more robust branching.

2496350 of our 10 million formula cells used an IF function,  
94978 cells contained a SUMIF (these numbers may overlap  
as a cell may have both a SUMIF and an IF (we noted over  
5000 unique formulas that had both an IF and a SUMIF)

The function breakdown between the three corpora is inter-  
esting. Some functions are used a similar amount (e.g. IF) and  
others are used differently (e.g. SUM, ISBLANK, VLOOKUP,  
ISTEXT). This underscores the need for large, diverse corpora.  
For example, all three corpora used IF about the same, whereas  
FUSE contains many more string-manipulation functions and  
Enron uses more financial functions. It may also be interesting  
to explore function "synonyms", which occur when there is  
more than one way to achieve the same result. For example,  
in FUSE and Enron, workbooks are more likely to use the +  
operator than the SUM function, but in EUSES, those tools  
are used at about the same rate.

A summary of your spreadsheet corpus can be found in  
Table III.

#### V. RESULTS

This dataset suggests many interesting questions.

TABLE III  
COMPARISON OF FUSE AND OTHER SPREADSHEET CORPORA

	FUSE	EUSES	Enron
Size (n)	249,376	6,000	15,570
Space (GB)	b	0.64	23.3
Research access	All	Researchers	All
Unique formulas	894361	693266	84004
Extendable	Yes	Not scalable	No
Framework	Hadoop	Excel + VBA	Scantool
Time Period	2006	2006	2006
Distinct Functions	219	209	139

TABLE IV  
SPREADSHEET CLASSIFICATION

Category	FUSE	EUSES
Database	4518	720
Finance	3058	780
Grade	2915	731
Homework	61	682
Inventory	2243	756
Model	2143	966

#### A. Classification of spreadsheets

The results in this section are intended to demonstrate the essential properties of the corpus.

B. RQ2: How diverse is the common common crawl corpus?

C. RQ1: How stable are URIs?

D. RQ3: NLP Extraction of Spreadsheet?

E. RQ4: What types of formulas are used by end-user software programmers?

Top two formulas are the same across the three corpora: 1. Add up the three cells to my left 2. Add up the two cells to my left And most of the top 10 are Add up n cells to my left. Fuse #3 HYPER-LINK("http://www.eia.doe.gov/totalenergy/data/monthly/dataunits.cfm,"Note: F. Hermans, and A. van Deursen, "Detecting code smells in spreadsheet formulas," in *Proceedings of the 2012 IEEE International Conference on Software Maintenance (ICSM)*, ser. ICSM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 409–418. [Online]. Available: <http://dx.doi.org/prox.lib.ncsu.edu/10.1109/ICSM.2012.6405300>) Enron #3 NOW()

give warc record ids

## VI. CHALLENGES AND LIMITATIONS

relevance is dependent on common crawl definition of relevant

## VII. CONCLUSION

The conclusion goes here.

## VIII. RELATED WORK

#### A. Why use spreadsheets

[14] Use spreadsheet corpus + interviews to determine which features end-users use. [7] Detecting code smells in spreadsheets. Analyze EUSES to study occurrence of smells. [8] Refactor spreadsheet formula. Perform case study using EUSES dataset. [2] Support debugging spreadsheets

#### B. What other corpora?

EUSES [12]

[11] Automatically extract relational data from spreadsheets. Extracted 410,554 spreadsheets from clue09 web crawl.

[13] ENRON find citation ;– icse seip 2015

[11] Automatically extract relational data from spreadsheets. Extracted 410,554 spreadsheets from clue09 web crawl.

**Existing corpora. Spreadsheet tools.**

## ACKNOWLEDGMENT

This material is based upon work supported in whole or in part with funding from the Laboratory for Analytic Sciences (LAS). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the LAS and/or any agency or entity of the United States Government.

## REFERENCES

- [1] M. Burnett, "What Is End-User Software Engineering and Why Does It Matter?" in *End-User Development SE - 2*, ser. Lecture Notes in Computer Science, V. Pipek, M. Rosson, B. de Ruyter, and V. Wulf, Eds. Springer Berlin Heidelberg, 2009, vol. 5435, pp. 15–28. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-00427-8\\_2](http://dx.doi.org/10.1007/978-3-642-00427-8_2)
- [2] R. Abraham, M. Burnett, and M. Erwig, "Spreadsheet Programming," in *Encyclopedia of Computer Science and Engineering*. John Wiley & Sons, Inc., 2009, pp. 2804–2810.
- [3] A. Blackwell, "First steps in programming: a rationale for attention investment models," in *Proceedings IEEE 2002 Symposia on Human Centric Computing Languages and Environments*, 2002, pp. 2–10.
- [4] A. J. Ko, B. Myers, M. B. Rosson, G. Rothermel, M. Shaw, S. Wiedenbeck, R. Abraham, L. Beckwith, A. Blackwell, M. Burnett, M. Erwig, C. Scaffidi, J. Lawrance, and H. Lieberman, "The state of the art in end-user software engineering," *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–44, Apr. 2011.
- [5] B. A. Nardi and J. R. Miller, "The spreadsheet interface: A basis for end user programming," in *Human-Computer Interaction: INTERACT '90*, 1990, pp. 977–983.
- [6] C. Scaffidi, M. Shaw, and B. Myers, "Estimating the numbers of end users and end user programmers," in *Visual Languages and Human-Centric Computing, 2005 IEEE Symposium on*. IEEE, 2005, pp. 207–214.
- [7] M. Erwig, F. Hermans, and A. van Deursen, "Detecting code smells in spreadsheet formulas," in *Proceedings of the 2012 IEEE International Conference on Software Maintenance (ICSM)*, ser. ICSM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 409–418. [Online]. Available: <http://dx.doi.org/prox.lib.ncsu.edu/10.1109/ICSM.2012.6405300>
- [8] S. Badame and D. Dig, "Refactoring meets spreadsheet formulas," in *Proceedings of the 2012 IEEE International Conference on Software Maintenance (ICSM)*, ser. ICSM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 399–409. [Online]. Available: <http://dx.doi.org/prox.lib.ncsu.edu/10.1109/ICSM.2012.6405299>
- [9] S. G. Powell, K. R. Baker, and B. Lawson, "A critical review of the literature on spreadsheet errors," *Decis. Support Syst.*, vol. 46, no. 1, pp. 128–138, Dec. 2008. [Online]. Available: <http://dx.doi.org/10.1016/j.dss.2008.06.001>
- [10] R. R. Panko, "Two corpses of spreadsheet errors," in *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*. IEEE, 2000, pp. 8–pp.
- [11] Z. Chen, M. Cafarella, J. Chen, D. Prevo, and J. Zhuang, "Senbazuru: A prototype spreadsheet database management system," *Proceedings of the VLDB Endowment*, vol. 6, no. 12, pp. 1202–1205, 2013.
- [12] M. Fisher and G. Rothermel, "The EUSES spreadsheet corpus," in *ACM SIGSOFT Software Engineering Notes*, vol. 30, no. 4. ACM, Jul. 2005, p. 1.
- [13] F. Hermans and E. Murphy-Hill, "Enrons spreadsheets and related emails: A dataset and analysis," 2015, to appear.

- [14] C. Chambers and C. Scaffidi, "Struggling to excel: A field study of challenges faced by spreadsheet users," in *Proceedings of the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, ser. VLHCC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 187–194. [Online]. Available: <http://dx.doi.org/prox.lib.ncsu.edu/10.1109/VLHCC.2010.33>