

Accounting Standard Citations in SEC Filings, 2000-2026

Data and Code

Jillian Bommarito^{1,2} Michael J. Bommarito II^{1,*}

¹Bommarito Consulting, LLC

²Department of Accounting and Information Systems, Eli Broad College of Business, Michigan State University

September 28, 2026

Abstract

Researchers often infer which accounting standards apply to a company from the length of the standards or the company’s XBRL tags. This paper provides a direct measure based on the standards that companies name in their filings. The dataset covers accounting-standard references in 10-K, 10-Q, 20-F, and 40-F filings on EDGAR for fiscal years 2000-2026. It links each resolved reference, including references to standards superseded in 2009, to a FASB Accounting Standards Codification topic. Five topics account for half of all citations. An update’s number is cited most in the year before it takes effect, when advance disclosure is required; the topic it amends peaks in the year it applies. Topics with more guidance also appear in filings from more companies. The release includes documented CSV tables by topic, year, industry, and form; a filer-year table that joins to Compustat by CIK; complete extraction code; and validation results.

Keywords: accounting standards, FASB Accounting Standards Codification, SEC filings, EDGAR, textual analysis, citation extraction, accounting complexity, open source code

JEL classification: M41, M48, C55, C81, C88

Data Availability: Data are available for download under a CC-BY licence from the deposit cited in the text (DOI pending) and from the Journal of Information Systems dataset repository. Code is available at <https://github.com/mjbommar/asc-filing-data>.

I. INTRODUCTION

Researchers cannot directly observe which parts of U.S. Generally Accepted Accounting Principles (GAAP) apply to each public company. No public register links every filer to the standards that govern its accounting. Prior studies therefore use indirect measures. They weight a company’s exposure by the length of applicable guidance (Donelson et al. 2012; Chychyla et al. 2019), count elements tagged with eXtensible Business Reporting Language (XBRL) (Hoitash and Hoitash 2018, 2022), or measure how widely the guidance was dispersed before the Codification (Binz et al. 2023). Yet companies often identify applicable standards in their own filings. Those filings are public and machine-readable, but no dataset has collected these references across the full filing population.

This paper provides that dataset. It covers all 10-K, 10-Q, 20-F, and 40-F filings on EDGAR, the SEC’s electronic filing system, for fiscal years 2000 through 2026, a total of 1,003,504 filings. The extraction finds 20,435,015 mentions of named accounting standards. After removing overlapping matches and

*Corresponding author: michael@bommaritollc.com. The authors used large language models, including Anthropic’s Claude and OpenAI’s Codex, to draft and revise text, write and review code, label validation samples under written rubrics, and conduct simulated editorial reviews. The authors reviewed the resulting text, code, and labels and take full responsibility for the published article.

duplicate documents, 19,535,536 references remain. The method links 96.3% of them to a topic in the FASB Accounting Standards Codification.

The link is direct when a filing names a Codification topic. When it names an Accounting Standards Update (ASU), the method uses the update calendar to find the topics that the ASU amended. Older filings present a harder problem because they cite pronouncements that the Codification replaced in 2009, including Statements of Financial Accounting Standards (SFAS), FASB Interpretations (FIN), Accounting Principles Board Opinions (APB), Accounting Research Bulletins (ARB), and Emerging Issues Task Force (EITF) issues. We link these pronouncements to their successor topics with transition sentences in which filers name both the old standard and the new topic. A context classifier then separates references that describe a company's accounting from references in an ASU title or a disclosure about a standard that is not yet effective. We call the first group substantive mentions.

The release contains documented CSV tables, the code that rebuilds them from EDGAR, and the results of the validation tests. Its main table reports the number of filers, filings, and mentions for each Codification topic and fiscal year. It also reports the population of periodic filers and the corresponding citation rates, both for substantive mentions and for all mentions. Other tables divide the results by two-digit Standard Industrial Classification (SIC) industry, state of incorporation, country of domicile, exchange, form, and filer type. The release also records first citations by filer, citations to each ASU relative to its issue and effective dates, and every distinct cited standard with its resolution. A filer-topic-year table supports joins to Compustat and the Center for Research in Security Prices (CRSP).

The dataset contributes a direct measure to four areas of accounting information systems research: financial reporting and systems, data and text mining, XBRL, and data analytics. It follows earlier *Journal of Information Systems* (JIS) releases of 10-K text sections (Codesso et al. 2026), XBRL element counts (Hoitash and Hoitash 2022), and SEC comment letters (Olsen and Hassan 2026). Those resources measure filing text, tagging, or regulatory review. The new dataset measures which accounting rules companies name. For the post-2009 period, researchers can compare what a filer tags in XBRL with what it discusses in prose. Standard setters can trace discussion of a rule through adoption and later use, while auditors can compare a client's disclosures with those of other filers subject to the same guidance.

Three design choices affect how users should interpret the measures. First, each annual rate uses the number of registrants that filed a periodic report that year as its denominator. That population fell 45 percent between fiscal 2000 and 2025. A fixed number of citing companies thus represents a larger share in later years. Second, the main rate includes only substantive mentions, while each table also reports the unfiltered rate. The two measures can differ by more than ten percentage points when a large update has been issued but is not yet effective. Third, the pre-2009 bridge comes from votes across filing text rather than a hand-coded crosswalk. Each mapping ships with its vote count, runner-up, and runner-up share so users can inspect or replace the choice.

The descriptive results help define what the measure captures. Citation is concentrated: five topics account for half of all citations, while the least-cited half of the Codification accounts for less than 1 percent. References to an ASU number peak in the year before the standard takes effect, when SEC Staff Accounting Bulletin No. 74 (SAB 74) requires disclosure of its expected effect. References to the amended topic peak in the year the standard applies and then fall as the standard becomes routine. Among the 90 topics in force throughout 2020–2024, the share of filers citing a topic has a Spearman correlation of +0.768 with the number of guidance paragraphs in that topic. This association supports the topic-level length weights used in prior complexity research, but it does not establish the same relation within companies.

Section II compares the dataset with prior measures and shared EDGAR resources. Section III documents the source data, each processing step, and the validation tests. Section IV describes the released files, reports descriptive statistics, and demonstrates one use of the data. Section V presents research questions that the data can support. The appendices provide the data dictionary, reproduction instructions, and merge details.

II. BACKGROUND AND RELATED LITERATURE

The dataset connects three lines of research: measures of a company's exposure to GAAP, textual data drawn from SEC filings, and citation extraction from legal documents. This section reviews the work needed to place the measure and the released files. Broader reviews of textual analysis in accounting appear in Loughran and McDonald (2016) and Bochkay et al. (2023).

Measures of Exposure to GAAP

Researchers have built several proxies for the standards that govern a company. Each proxy observes something other than a company's own reference to a standard. Table 1 compares four established approaches with the measure released here.

Table 1: Measures of a company's exposure to GAAP

Measure	What it observes	Grain and period	Main limit
Guidance length (Donelson et al. 2012)	Words in the standard	Standard; static	Does not show whether a company applies the standard.
XBRL tag count (Hoitash and Hoitash 2018, 2022)	Elements tagged by a company	Company-year; 2009–	Does not identify the standard behind an element or cover pre-XBRL years.
XBRL tags mapped to topics (Chychyla et al. 2019)	Elements joined to the reference linkbase	Company-year; 2009–	Can miss disclosure-only standards and does not cover the pre-2009 period.
Guidance dispersion (Binz et al. 2023)	Sources within each Codification area	Area; time-invariant	Is fixed across companies and years.
This dataset	Standards named in filing text	Filing-topic-year; 2000–2026	Misses an applicable standard when the company does not name it.

Section IV.5 examines the limits of using an explicit reference as evidence that a standard applies.

Guidance length provides the earliest of these measures. Donelson et al. (2012) define a transaction's complexity from the length of the governing pronouncement. Chychyla et al. (2019) extend that idea to companies: they identify applicable Codification topics through XBRL tags and weight each topic by its volume of guidance. Other studies treat the number of XBRL elements as a measure in its own right (Hoitash and Hoitash 2018). The JIS dataset paper by Hoitash and Hoitash (2022) releases that company-year measure and its SAS code. At a different level, Binz et al. (2023) measure the dispersion of pre-Codification guidance within each area of GAAP. Restatements are corrections of prior financial reports; their study tests whether the 2009 consolidation reduced them. Researchers have used these measures to study restatements, audit fees, analyst forecast accuracy, and SEC comment letters.

A related literature measures the filing rather than the governing standards. Annual reports have become longer and harder to read (Li 2008; Loughran and McDonald 2014). Boilerplate and repeated language account for much of that growth (Dyer et al. 2017). Researchers also use disclosure length as a measure of reporting complexity (Peterson 2012; Filzen and Peterson 2015), although companies may

respond to complex financial statements with more voluntary disclosure (Guay et al. 2016). Scott et al. (2024) measure how much disclosure text comes from standard-setter guidance. Their measure captures copied language; ours captures an explicit reference to the standard. The two can be used together.

The context labels also connect the dataset to research on standards awaiting adoption. SEC Staff Accounting Bulletin No. 74 (SAB 74) asks a registrant to describe the expected effect of a standard that has been issued but is not yet effective. Prior studies find that these disclosures contain information about the standard's eventual effect (Alexander et al. 2011; Ahn et al. 2021). Our data identify references made in that setting separately from descriptions of accounting already in use. None of these prior measures directly observes a company's statement that a named standard governs its accounting.

Shared Data from EDGAR Text

Several projects distribute EDGAR text or measures derived from it. The Notre Dame Software Repository for Accounting and Finance provides cleaned 10-K and 10-Q filings and document dictionaries from 1993 onward (Loughran and McDonald 2016). That collection offers a natural corpus for an independent replication. We instead use the SEC dissemination feed because it also includes 20-F and 40-F filings by foreign private issuers. Lu et al. (2026) release a model that segments 10-K filings into items with a macro-F1 of 0.98. Their result provides a useful benchmark for extraction from filing text, although the present task extracts citations rather than document sections.

JIS has published several related datasets with their construction code. Codesso et al. (2026) extract major sections and financial-statement notes from 10-K and 10-Q filings beginning in 2008. Hoitash and Hoitash (2022) release a company-year XBRL complexity measure, and Bonrath et al. (2025) provide code for retrieving and categorising SEC filings. Other releases cover SEC comment letters (Olsen and Hassan 2026), cybersecurity risk disclosures (Calderon and Gao 2023), and a human-capital lexicon with its training disclosures (Demers et al. 2024). The same JIS track has also published Twitter data (Wang 2022), Form 990 data (Abu-Khadra and Olsen 2023), Glassdoor reviews (Zhou et al. 2024), and experimental data (George et al. 2023).

The present release complements these resources. It indexes the accounting standards cited in filing text from 2000 onward and links old pronouncements to the post-2009 Codification. Researchers can join it to other EDGAR resources by accession number or SEC Central Index Key (CIK). This design separates the source text from the new measurement layer and permits independent checks against another filing corpus.

The extraction method also draws on citation analysis outside accounting. Studies of the United States Code (Bommarito and Katz 2010) and statutory law across countries (Coupette et al. 2021) extract citations from a body of legal text, resolve them to the source hierarchy, and count the resulting links. This dataset applies the same sequence to financial reports that cite accounting standards. The main difference is temporal: the target system was renumbered in 2009, while more than half of the observed references use the older names.

Prior Uses of the Dataset

No published study has used the complete dataset. Two source components predate it. The Codification structure tables and the ASU calendar were first built for a public Codification reference site. A companion legal analysis examines public access to the standards indexed here (Bommarito II 2026). The citation extraction, pre-2009 bridge, context classifier, and released tables were created for this paper. Users should cite the paper for the construction method and the deposit DOI for the specific data release.

III. DATASET AND METHOD

The dataset comes from the SEC’s Electronic Data Gathering, Analysis, and Retrieval system (EDGAR). Five scripts build it in the sequence described below. Table 2 reports the sample at each stage.

Table 2: Sample selection

	Filings	Removed
Periodic reports in the corpus store (Forms 10-K, 10-Q, 20-F, 40-F, and variants), fiscal 1999–2026	1,025,205	
reports with a valid fiscal period year, 1999–2026 (the denominator population)	1,020,108	5,097
reports for fiscal 2000–2026 (fiscal 1999 is partial)	1,003,504	16,604
reports that name at least one accounting standard	782,160	221,344
citing reports in complete fiscal years, 2000–2025	771,410	10,750

Filings without a standard reference remain in each denominator. This table removes them only to report the number of filings summarised in the citation tables. The number of periodic filers ranges from 7,842 to 15,003 across fiscal 2000–2025 (denominator_by_year.csv).

Source and Sample

The source sample contains every periodic report in the SEC dissemination feed with a fiscal period from 1999 through 2026. It includes Forms 10-K and 10-Q, Forms 20-F and 40-F filed by foreign private issuers, and amendments and transition variants of those forms. Matching form names by prefix captures variants such as 10-K/A, 10-KT, and 10-K405.

The small-business forms require a separate rule. Their names, 10KSB and 10QSB, omit the hyphen and do not match the 10-K or 10-Q prefixes. Excluding them would remove 144,248 filings, or 14.1 percent of the corpus, all before 2009. Because small-business registrants are smaller and disclose less than other filers, that omission would also shift the early topic distribution toward large companies.

Fiscal 1999 is incomplete because the source store begins during that year. Fiscal 2026 is incomplete because the year is still in progress. The released tables retain both years and mark them as partial so users can see the limits of the series. The statistics reported in this paper exclude both years.

Step 1: Building the Corpus Store

The script `ingest_full.sh` runs the authors’ ingestion tool over the EDGAR full-index and daily feeds (Appendix B) and downloads each sampled filing. It extracts the primary document and stores its text in a compressed, content-addressed archive keyed by accession number. Each record also carries the filing header: form, filer CIK, period of report, and filing date. The resulting store is 77 GB. It is not part of the data release because the script reproduces it from the public EDGAR feed in about 12 hours.

The analysis assigns each filing to the fiscal year of its period of report, not its filing year. This choice aligns company references with the effective dates of standards. Users joining the data to a calendar-year source must convert the year field. The dissemination feed leaves the filing header’s SIC field blank, so industry comes from the SEC filer catalogue and is treated as a filer attribute.

Step 2: Extracting References

The script `extract_and_build.sh` makes two passes through the corpus. The first pass writes one record for every reference to a specific standard. The record includes the surrounding sentence and paragraph, the filing's Item and Part, the nearest heading, and an indicator for references inside tables.

The extractor recognises 11 citation schemes: Accounting Standards Codification (ASC) topics and subtopics, spelled-out Codification references, ASU numbers, U.S. Financial Accounting Standards Board (FASB) Staff Accounting Bulletins, and the pre-2009 SFAS, FIN, EITF, APB, and ARB series. The case-insensitive patterns cover forms observed in the corpus, including “ASC 842”, “Subtopic 842-20”, “FASB Statement No. 13”, and “Accounting Principles Board Opinion No. 25”. The scripts and Supplemental Materials contain the complete term list. A three-digit number counts as a topic only when that topic exists in the Codification. This check rejects strings such as “Topic 500” and stray page numbers.

The first pass finds 20,435,015 mentions. It then removes overlapping character matches within a document and identical text filed as both an original and an amendment, leaving 19,535,536 mentions. The second pass counts generic references to “GAAP” and “the FASB” at the document level. Their 14,932,381 individual occurrences do not carry distinct standard information, while the count per document does. The script tracks “non-GAAP” separately so it does not enter the GAAP count. Extraction takes less than one hour. Users can extend the pipeline by adding citation forms to the term list.

Step 3: Resolving References to Topics

Table 3: References linked to a Codification topic, by citation scheme

Scheme	Mentions	Distinct references	Linked (%)
ASC topic / subtopic	5 049 855	1 634	100.0
Accounting Standards Update	3 044 675	471	98.5
SFAS (FASB Statement)	8 220 990	211	98.5
FIN (FASB Interpretation)	1 107 381	65	98.2
APB Opinion	562 913	79	99.3
ARB	107 225	30	99.1
EITF Issue	547 047	316	79.1
SEC Staff Accounting Bulletin	492 580	122	98.0
FASB Staff Position (not linked)	402 870	155	0.0
All	19 535 536	3 083	96.3

Mentions are linked when they resolve to a Codification topic. Pre-2009 references link through the corpus-voted bridge (`data/reference/legacy_bridge_corpus.csv`); FASB Staff Positions are their own scheme and link to no topic by design. Largest unlinked legacy references: FSP SFAS-157 (71,946), FSP EITF-03-06 (38,639), FSP APB-14 (36,604), FSP SFAS-115 (34,665).

The script `build_model.py` links each mention to a Codification topic according to its citation scheme. Table 3 reports the resolution rate for each scheme. A Codification reference names its topic directly. An ASU reference links through the update calendar to every topic that the update amended. Among the 234 updates issued through 2026, 206 amend at least one Codification topic. The remaining updates concern SEC content or technical matters with no topic link. 25 updates amend more than one topic. Each amended topic receives one mention, and the ASU table reports the number of linked topics so users can apply fractional weights instead.

Staff Accounting Bulletins use a hand-coded map of the 17 bulletins that contain accounting guidance rather than disclosure guidance. These bulletins are too few, and too rarely cited with a topic, to support the voting method used for older standards. The release marks this map as the only reference table not built from the filing corpus.

References to pre-2009 standards require a bridge from each superseded pronouncement to the topic that absorbed it. The former FASB cross-reference table is no longer available on the Codification website. Review also found errors in a hand-coded copy of the archived table, including three EITF revenue issues linked to ASC 606 five years before that topic existed. We therefore build the bridge from the filing corpus with `build_legacy_bridge.py`.

A sentence that names an old pronouncement and a Codification topic casts one vote for that link. This pattern is common in transition disclosures, such as a sentence stating that SFAS 157 is now ASC 820. Under this rule, SFAS 157 maps to ASC 820 by 7,449 votes against 73 for the runner-up. A bridge row is confident when it has at least 20 votes from five filers and one topic receives at least 60 percent of them. Of the 245 legacy references co-cited with a topic, 104 meet that rule. Those rows account for 92.4 percent of all legacy mentions.

The archived hand table contains 53 entries, including 40 confident bridge rows. The corpus vote agrees with 36 of the 40. The four differences trace to errors in the hand table: three EITF revenue issues linked too early to ASC 606, and EITF 96-18, for which the corpus-supported mapping remained. Reading the underlying pronouncements led us to override three votes: APB 28, SFAS 48, and SFAS 45. In each case, a Staff Position or an unrelated co-citation drove the vote. The resolver records the reason for each override.

The bridge assigns a pronouncement absorbed into several topics to the plurality topic. The released transition sentences allow users to construct a one-to-many mapping instead. Citations to FASB Staff Positions and citations inside ASU titles do not cast votes. Across legacy mentions, 92.4 percent use confident voted rows, 0.7 percent use the three reviewed overrides, and 1.8 percent use archived-table rows below the confidence threshold. Another 2.7 percent refer to 15 amendment standards, including SFAS 145, 149, 151, and 153, that filers cite without naming the amended topic. These links were read from the pronouncements and are marked as hand-coded. The remaining 2.4 percent stay unresolved: they count as mentions but do not enter a topic rate.

Resolution reaches 98.5% for FASB Statements, 79.1% for EITF issues, and 96.3% across all schemes. Before 2009, topics are visible only through pronouncements linked by the bridge. The release therefore includes the vote count, runner-up, and runner-up share for every bridge row.

Step 4: Classifying Context and Dating Updates

A cited topic can enter a filing for several reasons. An ASU title follows the form “*Subject* (Topic NNN)”, so listing ASU 2023-07, “Segment Reporting (Topic 280)”, creates a Topic 280 reference. A disclosure under SEC Staff Accounting Bulletin No. 74, abbreviated SAB 74, cites an issued standard before it takes effect. Other references describe the company’s accounting already in use. These settings call for separate measures.

The rule-based classifier reads the sentence and paragraph around each mention. It applies a written rubric in fixed order: stated adoption first, not-yet-effective language second, and a topic number within a quoted title last. Each mention receives one of three labels: substantive, not_yet_adopted, or asu_title. Every released rate appears both for substantive mentions and for all mentions. The distinction can be large. In fiscal 2023, when ASU 2023-07 had been issued but was not yet effective, 70 percent of the 13,362 Topic 280 mentions appeared in not-yet-effective disclosures. By share of filers, Topic 280 ranked 3 on all mentions and 13 on substantive mentions.

The script `build_asu_effective_corpus.py` also extracts the effective dates that filers state in SAB 74 paragraphs. Dates receive votes across the corpus. When filers report different years because of a deferral or a later date for smaller reporting companies, the ASU record keeps the leading date and its runner-up.

Two documented deferrals of the public-company date (ASU 2014-09 and ASU 2018-12) take the later date, flagged. The record also stores the first fiscal year beginning after the stated date, which is the year the update first applies. Provenance files for both corpus-built references—the legacy bridge and the effective-date calendar—identify the extract used to produce them.

Step 5: Aggregating the Released Tables

The final step counts mentions, filings, and distinct filers for each topic and fiscal year. It divides the filer count by the number of registrants that filed a periodic report for that fiscal year. The same run writes the main table, cross-tabulations, first-use table, ASU table, and combined series of legacy and Codification references.

The denominator comes from the corpus because the filer population changed substantially. It fell from 14,685 for fiscal 2000 to 8,062 for fiscal 2025, a 45 percent decline. A fixed denominator would distort the trend. Aggregation takes about 20 minutes. A flag produces the operating sample; every figure and table in this paper uses the full sample.

Identifiers and Merging with Compustat and CRSP

Every filing record includes the SEC Central Index Key (CIK) and period of report. The released `asc_filer_topic_year` table reports mentions, substantive mentions, filings, and first-use year for each CIK, fiscal year, and topic. Users can join it to Compustat by matching CIK and the fiscal year of `DATADATE`. Compustat's company file supplies the `gvkey` for each CIK, and the CRSP/Compustat Merged link table maps `gvkey` to `permno`. Appendix B gives the join code. Audit Analytics and the SEC Financial Statement Data Sets also use CIK directly. The release contains no data from those products.

The year variable again represents the fiscal period, not the filing date. Industry, exchange, and state of incorporation come from the SEC filer catalogue. Because the catalogue supplies current attributes, a company that changed industry carries its current SIC code throughout the panel.

Validation

Validation addresses three questions: whether the extractor finds references present in the source text, whether extracted strings are valid references, and whether the context classifier assigns the correct label. The samples span filing eras, forms, and citation schemes. Each test exposed an error or a limit that changed the release.

Recall. Table 4 reports recall by era and form family. The first test used a seeded, stratified random sample of 423 filings read from the corpus store. A detector with broader Codification-era patterns scanned the same text. The extractor missed few references after 2015 but missed 6–14 percent of distinct references per filing before 2009. Most misses were FASB Statements written in unrecognised forms or phrases split by hard line breaks in plain-text filings. Because an era-specific recall gap could create a false shift at the 2009 Codification, the sample was stratified by era.

We added every observed surface form and reran the extraction. The amended extractor's score on that first sample is in-sample. We therefore drew a new sample of 210 filings with a different seed and removed every filing used in the first test. On this holdout, the extractor finds 99.5 percent of distinct references and at least 96 percent in every cell.

The recall estimate has two limits. For pre-2009 schemes, the comparison detector shares the extractor's patterns and thus tests only the forms already represented in the term list. A broader detector also accepts bare forms such as "Statement No. 151", "Interpretation No. 46", and "Opinion No. 25". Against that detector, recall on the same filings is 97.6 percent, with an interval from 96.7 to 98.3. The lowest cell is the 2003–2008 10-Q cell at 90.7 percent. A small era gradient therefore remains in the earliest quarterly reports. Adding these bare forms is a planned extraction change. The test also begins after ingestion, so it cannot measure

Table 4: Reference-extraction recall by period and form

Period years	Forms	Filings	Refs	Before		After		Held out		
				Recall (%)	95%	Recall (%)	95%	Filings	Recall (%)	95%
2000–2002	10-K	30	169	88.8	[83.0, 93.1]	100.0	[97.8, 100.0]	15	100.0	[95.5, 100.0]
2000–2002	10-Q	30	100	88.0	[80.0, 93.6]	100.0	[96.4, 100.0]	15	100.0	[91.0, 100.0]
2000–2002	20-F / 40-F	30	344	92.4	[89.1, 95.0]	100.0	[98.9, 100.0]	15	99.5	[97.0, 100.0]
2000–2002	10-KSB / 10-QSB	30	74	87.8	[78.2, 94.3]	100.0	[95.1, 100.0]	15	100.0	[86.8, 100.0]
2003–2008	10-K	30	276	92.0	[88.2, 94.9]	99.3	[97.4, 99.9]	15	100.0	[98.3, 100.0]
2003–2008	10-Q	30	179	87.2	[81.3, 91.7]	98.3	[95.2, 99.7]	15	100.0	[96.9, 100.0]
2003–2008	20-F / 40-F	30	398	93.7	[90.9, 95.9]	99.5	[98.2, 99.9]	15	100.0	[98.0, 100.0]
2003–2008	10-KSB / 10-QSB	30	125	86.4	[79.1, 91.9]	99.2	[95.6, 100.0]	15	100.0	[95.1, 100.0]
2009–2014	10-K	30	172	93.6	[88.8, 96.8]	97.1	[93.3, 99.1]	15	97.1	[92.7, 99.2]
2009–2014	10-Q	30	255	92.5	[88.6, 95.5]	98.8	[96.6, 99.8]	15	99.1	[95.3, 100.0]
2009–2014	20-F / 40-F	30	194	94.8	[90.7, 97.5]	99.0	[96.3, 99.9]	15	96.0	[86.3, 99.5]
2009–2014	10-KSB / 10-QSB	3	59	94.9	[85.9, 98.9]	100.0	[93.9, 100.0]			
2015–2026	10-K	30	261	98.5	[96.1, 99.6]	98.5	[96.1, 99.6]	15	100.0	[96.7, 100.0]
2015–2026	10-Q	30	217	100.0	[98.3, 100.0]	100.0	[98.3, 100.0]	15	100.0	[96.8, 100.0]
2015–2026	20-F / 40-F	30	131	99.2	[95.8, 100.0]	99.2	[95.8, 100.0]	15	100.0	[93.6, 100.0]
All		423	2954	93.2	[92.2, 94.1]	99.2	[98.8, 99.5]	210	99.5	[99.0, 99.8]

Before: thirty filings per cell (three in the 2009–2014 small-business cell, all there are), drawn with a fixed seed from filings over 20,000 characters, read back from the store and scanned with a detector broader than the extractor for Codification-era surfaces; a reference counts once per filing; intervals are exact at the reference level and understate filing-level uncertainty. *After*: the same filings against the re-extracted corpus, an in-sample figure. *Held out*: fifteen per cell with a new seed after the amendment, first-sample filings removed (none remain for 2009–2014 small business); 210 filings, 1498 references; for pre-2009 schemes the detector shares the extractor’s patterns, and a broader detector gives 97.6 percent on the same filings.

Table 5: Context classifications in the held-out labelled sample

Labelled	Classified as			Total
	Substantive	Not yet adopted	ASU title	
Substantive	99	6	2	107
Not yet adopted	4	35	3	42
ASU title	1	0	0	1
Precision	0.95	0.85	0.00	
Recall	0.93	0.83	0.00	

The holdout contains 150 of 450 mentions sampled by scheme and era. Large-language-model readers (Claude Opus) applied the written rubric in qa/context_label_rubric.md in three batches. The other 300 labels set the classifier’s precedence rules. Overall accuracy is 89.3 percent (83.3–93.8). The classifier reports a substantive share of 69.3 percent, compared with 71.3 percent in the labels.

source documents lost during ingestion. That limit matters because the share of filers naming any standard falls from 93 percent in fiscal 2009 to 76 percent in 2013. We therefore fetched 100 filings from 2011 to 2013 that name no standard in the store directly from EDGAR. In 99 of them, the primary document names no standard either. The decline reflects what filers wrote, not ingestion loss. The release includes both samples, both detectors, and both seeds.

Precision. An initial review of 300 matches from 2015–2024 found two false positives caused by a drug name truncated to “ARB 14”; the pattern was corrected. Because this test preceded both re-extraction and the addition of pre-2009 forms, we drew a second seeded sample of 548 matches from the final corpus. The sample contains up to 12 matches from each of 49 era-by-scheme cells, so every released scheme appears. A model reader applied a written labelling rule. Strict precision—whether the surface is the cited identifier—is 96.0 percent, with an interval from 94.0 to 97.5.

19 of the 22 false positives arise when a FASB Staff Position such as “FSP APB 14-1” is matched as “APB 14”. The release therefore records a Staff Position as its own scheme and links it to no topic. The rule flags a surface only when the Staff Position identifier stands immediately before it or a suffix such as “-1” follows it. It also flags a bare parent reference in a paragraph that names FSP FAS 107-1 and APB 28-1 or FSP FAS 117-1. The rule flags 402,870 mentions, clustered in legacy schemes from fiscal 2003 through 2014. Before this change, FSP FAS 107-1 and APB 28-1 had raised Topic 270 to 47 percent of filers in fiscal 2009; it is now 5 percent, against 2 percent in 2008 and 1 percent in 2010.

The other three false positives are a company named “FIN2”, a page number following “SAB”, and a retirement plan named “Hot Topic 401(k)”. The resolver rejects all three identifiers, so they do not change a topic rate. The release includes the sample, judgements, and labelling rule.

Context classification. Table 5 compares the classifier with 450 labelled Codification-era mentions. The sample covers ASC, ASU, and Topic references from fiscal 2008 through 2026 and is stratified by scheme and era. Large-language-model readers (Claude Opus) applied the written rubric in three batches. The first 300 labels set the precedence rules, while 150 were held out. Two further strata, described below, cover the not-yet-adopted class and pre-2009 references.

The first classifier scored 75.8 percent on all 450 labels and understated the substantive share by 12 percentage points. It tested for a topic number inside a quoted title before testing for adoption language. Reversing that order, then testing SAB 74 language and treating title references as the residual, raised holdout accuracy to 89 percent. The resulting substantive share is within two percentage points of the labelled share. A second model reader labelled 100 of the 450 examples from the rubric alone. Raw agreement was 95 percent, with Cohen’s κ of 0.88.

The ASU-title class was too rare in the stratified sample to estimate its precision. We therefore labelled 120 mentions that the classifier assigned to that class. Only 35 percent were ASU-title references, so the released column should be read as an upper bound on that mechanism.

Two further strata were labelled after the first review. The first holds 150 ASU mentions from fiscal 2014–2019, drawn 100 to 50 on the classifier’s own prediction. Accuracy is 84 percent (77 to 90). Reweighted to the population, where 60 percent of such mentions are called not-yet-adopted, recall for that class is 0.85. About 15 percent of SAB 74 mentions therefore enter the substantive series in those years. The second holds 150 pre-2009 legacy mentions drawn uniformly. Accuracy is 82 percent (75 to 88), not-yet-adopted recall is 0.77, and the substantive share is overstated by +3.4 points.

Three further limits apply. First, the adoption rule reads the full paragraph and matches “upon adoption”. It can label a future standard as substantive when the same paragraph describes another adoption or uses future-tense SAB 74 language. Half of the labelled mentions captured only by that phrase were not yet adopted. The substantive series may therefore lead the true effective date, especially at the start of the anticipation window. Second, both sets of labels came from model readers. Their agreement does not replace a human-labelled subsample for research that depends on these context classes. Third, the rubric

reflects a measurement choice. The released contexts, rubric, labels, and sampler allow users to apply a different rule.

IV. DATA DESCRIPTION AND STATISTICS

Released Files

Table 6 lists the released files. Each table is a CSV file, and Appendix A defines every field. The main table, `asc_topic_year`, has one row for each topic and fiscal year. It reports mention, filing, and filer counts; the annual filer denominator; substantive and unfiltered citation rates; the topic's annual rank; and a flag for years with incomplete form coverage. The cross-tabulations use the same fields and a denominator calculated within each group.

Three tables support questions that the main panel cannot answer. `asc_topic_first_use` records the year of each filer's first substantive reference to a topic. Unlike annual mention counts, first use cannot later decline for the same filer. It is left-censored at the start of the panel, and a new registrant's first filing counts as a first use; an incumbent-only count excludes filers new to the panel. `asu_citation_year` places ASU citations on two event clocks: years since issue and years since effective date. `topic_continuous` combines references to legacy standards with references to their successor topics, producing a series that crosses the 2009 change in names. The filer-level table, `asc_filer_topic_year`, includes CIK and provides the observation grain for company-data joins.

The main panel and its principal cross-tabulations ship for two samples. The full sample contains every registrant. The operating sample excludes funds, blank-check companies, structured-finance trusts, and filers without an SIC code. Its denominator is recalculated each year after those exclusions. The operating sample is therefore the appropriate version for claims about the share of operating companies that name a standard.

Coverage by Year

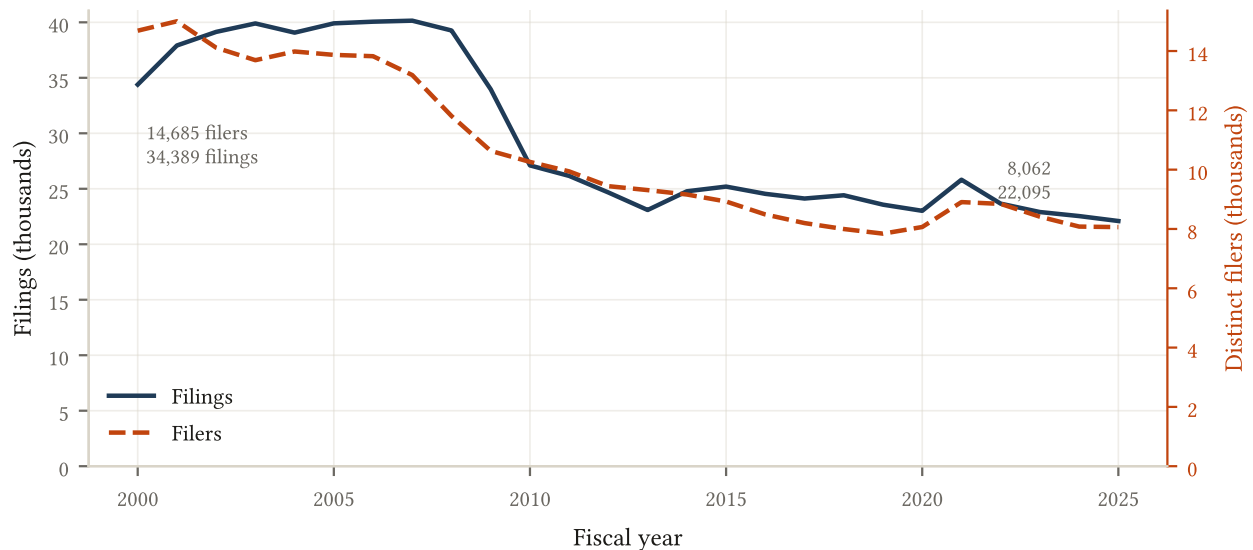


Figure 1: Registrants filing periodic reports (all of them), and filings with at least one standard reference, 2000–2025. The number of filers falls 45 percent, while filings with a reference fall 36 percent and all filings fall 56 percent. Annual citation rates use the filer count as their denominator.

Table 6: Released files

File (.csv)	Grain	Rows	Contents
asc_filer_topic_year	filer $\times$ fiscal year $\times$ topic	1,762,495	Mention and filing counts and first-use year for each CIK; used for company-level joins (supplement and deposit)
asc_topic_year	topic $\times$ fiscal year	2,051	Main panel with counts, denominator, substantive and unfiltered rates, rank, and a complete-year flag
asc_topic_sic	topic $\times$ year $\times$ 2-digit SIC	69,425	Main panel fields with a denominator for each SIC-year group
asc_topic_state	topic $\times$ year $\times$ state	41,992	State of incorporation
asc_topic_domicile	topic $\times$ year $\times$ country	9,586	Country of domicile
asc_topic_exchange	topic $\times$ year $\times$ exchange	7,191	Listing exchange
asc_topic_form	topic $\times$ year $\times$ form	6,779	Form family: 10-K, 10-Q, 20-F, 40-F, or small-business form
asc_topic_filertype	topic $\times$ year $\times$ filer type	3,893	Operating company, fund, blank-check vehicle, trust
asc_topic_first_use	topic $\times$ fiscal year	1,850	Number of filers that first name the topic in that year
asu_citation_year	ASU $\times$ fiscal year	2,767	References to each update relative to its issue and effective years
asc_legacy_year	legacy standard $\times$ year	6,892	References to pre-2009 pronouncements and their linked topics
topic_continuous	topic $\times$ year	921	Combined legacy and Codification references across the 2009 transition
codification_transition	fiscal year	13	Filers naming the Codification, a legacy standard, or both
asc_reference_inventory	distinct reference	3,083	Each standard identifier found in the corpus, its scheme, and its topic link
asc_topic_takeup	topic	63	Share of filers in the topic's primary two-digit SIC that name it
asc_topic_xbrl	topic $\times$ year, 2024–2026	162	Filers tagging an element whose only authoritative topic reference is this topic
asc_deadletter	topic	90	Topics named by less than 1 percent of filers, with industry and XBRL checks
corpus_coverage	fiscal year	28	Annual counts of filers, citing filers, and citing filings

Each file also has a JSON version, and Appendix A defines its fields. The directory `derived_operating/` repeats the topic-year, cross-tabulation, first-use, ASU, legacy, inventory, and coverage tables for operating companies. The file `sample_filers.csv` gives the reason for each exclusion. Because of its size, the compressed filer-level CSV appears in the deposit but not the code repository. The `reference/` directory contains the topic list, ASU calendar, corpus-voted bridge, voted effective dates, vote counts, and annual denominators.

Figure 1 shows the denominator for the annual rates. The number of registrants filing periodic reports falls from 14,685 in fiscal 2000 to 8,062 in fiscal 2025. A topic cited by the same number of companies in both years therefore has almost twice the later share. The decline is not steady. New registrants raise the denominator by about 10 percent in 2021 and 2022 before it falls again. About four in ten of those new filers are blank-check companies, which the operating sample removes.

Many filings do not name an accounting standard. From fiscal 2000 through 2025, 220,516 of 991,926 filings, or 22.2 percent, contain no matched reference. This share falls from 33.5 percent in fiscal 2000–2002 to 15.1 percent after 2014. Short reports and foreign-issuer forms account for much of the difference. About half of 20-F and 40-F filings name no U.S. standard because many foreign private issuers report under International Financial Reporting Standards (IFRS). These filings remain in the denominator. Restricting the denominator to citing filings would remove about one-fifth of the population and would place more weight on large domestic filers.

Annual topic coverage also changes at the 2009 Codification. Before 2009, companies named pronouncements rather than topics, so a topic enters the early panel only through a bridged pronouncement. The data contain 50 topics in 2003, 58 in 2008, and 86 in 2009, the first year in which filers could name topics directly. Topics supported by many dispersed sources rather than one numbered pronouncement are most likely to be absent from the early panel.

Differences across Topics

Figure 2 ranks topics by the share of filers citing them in fiscal 2022. We use 2022 because the 2023 segment and income-tax updates raised references to Topics 280 and 740 before those updates took effect. Income taxes rank first, followed by revenue, derivatives, leases, fair value, stock compensation. The order changes from year to year, but the leading group is stable. Between fiscal 2016 and 2025, the Spearman correlation between adjacent annual rankings ranges from 0.92 to 0.97. Income taxes, revenue, stock compensation, fair value, and leases appear in the top six every year from 2018 through 2025.

For the median topic, substantive and unfiltered citation rates differ by less than 0.25 percentage points. The difference is much larger during the period before a major standard takes effect. For ASC 326, current expected credit losses (CECL), the gap reaches 29 points in fiscal 2019, while the standard was pending: 47 percent of filers cite the topic, but only 18 percent describe it in substantive contexts.

References are concentrated among a small number of topics. Five topics account for 50.6% of all citations in 2020–2024, and 20 account for 91.3%. The least-cited half of the Codification accounts for 0.8%, producing a Gini coefficient of 0.830 across topic shares. Among the 90 topics in force throughout this period, the number of guidance paragraphs has a Spearman correlation of +0.768 with the share of filers citing the topic. This result is consistent with the length weighting in Donelson et al. (2012). It also reflects what filers choose to name: a widely applicable topic can remain in the tail when companies rarely discuss it by number.

Application: The Year Before a Standard Takes Effect

Figure 3 demonstrates one use of `asu_citation_year` without testing a hypothesis. It plots citation weight by years from an update's effective date, using dates extracted from filers' SAB 74 disclosures. Filers state the date as a boundary, such as fiscal years beginning after December 15, 2018, so the first year of application is the following fiscal year for a calendar-year company. Across 128 updates, 53.4% of the citation weight falls before that year. Weight peaks in the year before application. In the year after application, weight is 0.38 of its level in the year of application.

Some filers report different effective years for the same update because of deferrals or delayed application to smaller reporting companies. The main figure uses the most common year. This choice is ambiguous for 67 updates. Excluding those updates gives a pre-effective share of 54.9%. Using the earliest and latest

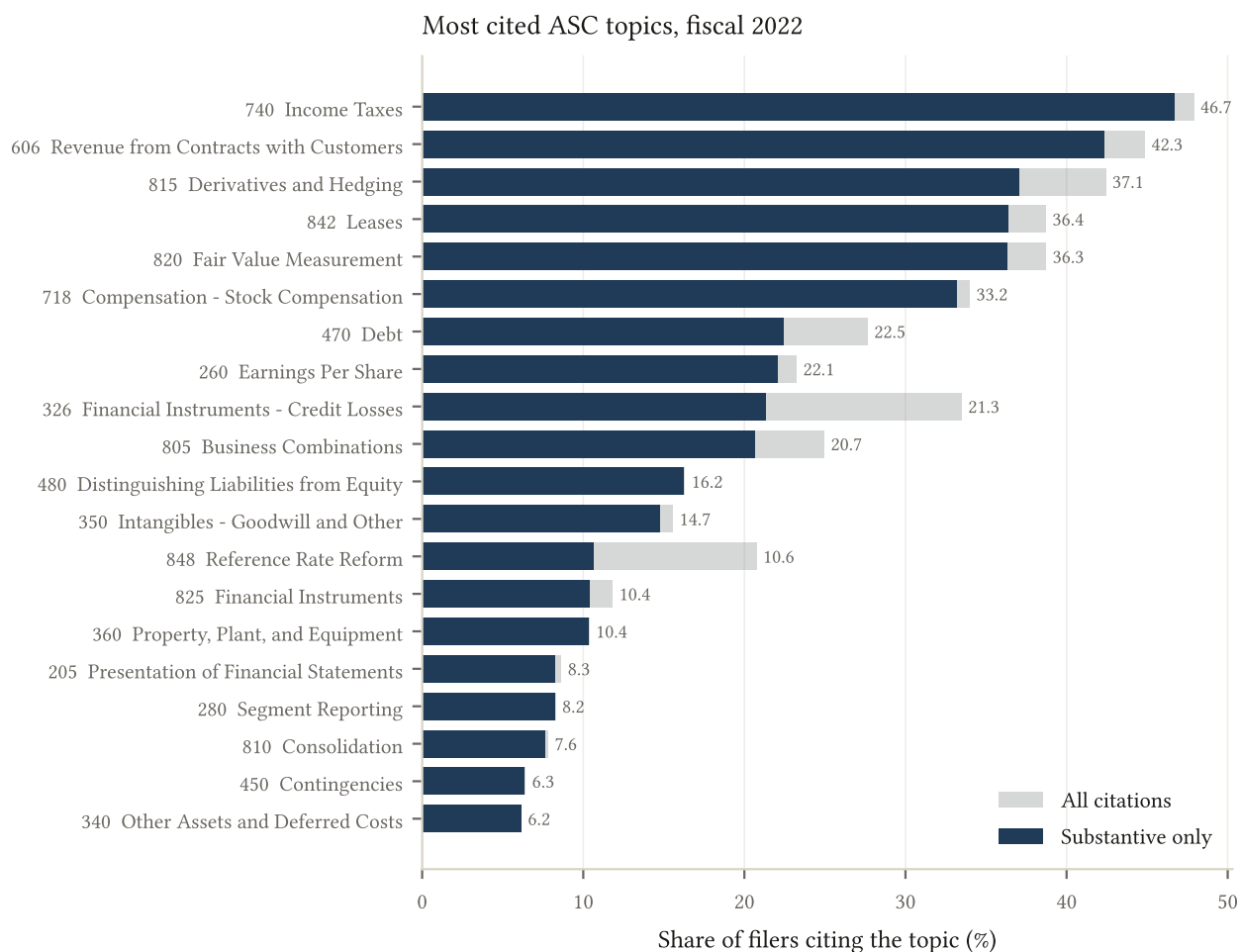


Figure 2: Most cited ASC topics in fiscal 2022. Pale bars include all references. Solid bars exclude references found only in an ASU title or a not-yet-effective disclosure. The gap measures how much reported prominence precedes application. Value labels give the substantive rate.

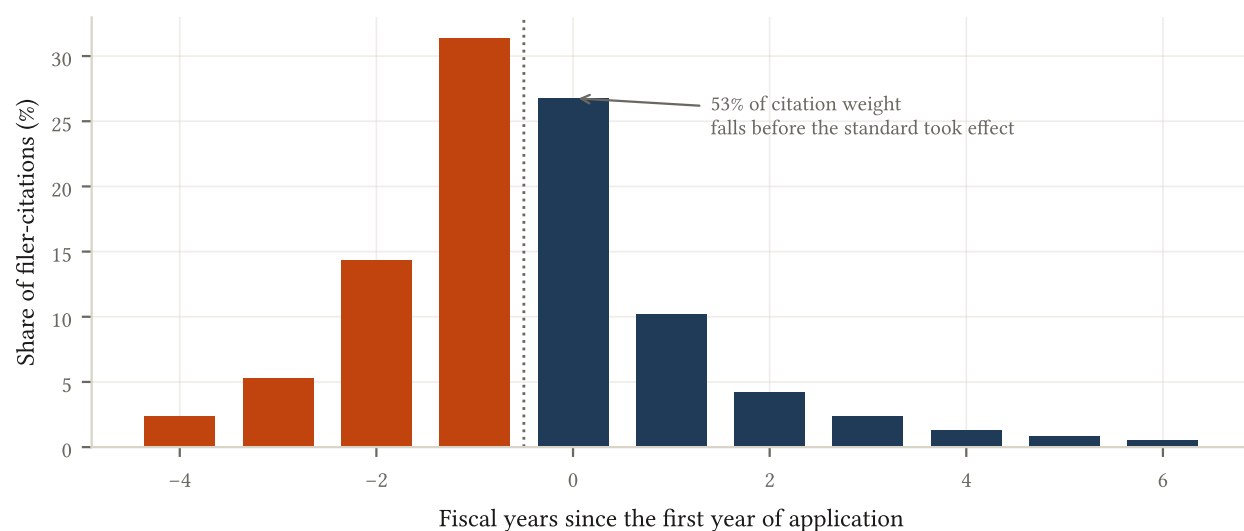


Figure 3: ASU citation weight by fiscal year relative to the first year of application for the 128 updates with an effective date stated by filers. Values before zero are references made before the update applied.

reported dates gives 51.4% and 62.3%, respectively. The peak remains in the year before application under every rule.

The pattern separates anticipation from application. More than half of the citation weight appears before the first year of application, mostly in the preceding year, when SAB 74 requires disclosure. Filers name a pending standard by its ASU number. Once it applies, they describe their accounting under the amended topic, and the ASU number falls away. The topic series show the other half: substantive references to ASC 606 peak in fiscal 2018 and to ASC 842 in 2019, each the first year of application. A measure that treats every ASU citation as adoption will date events a year early. Related transitions show the same overlap. In fiscal 2009, 69.0% of filers name both the Codification and a superseded standard. References to ASC 605 and ASC 606 also peak together in fiscal 2018 because transition notes name both the old and new guidance.

Advantages and Limitations

The dataset improves on the proxies in Table 1 in three ways. It observes a company's own reference to a standard, covers years before and after the 2009 Codification, and releases the inputs behind each measurement choice. Users can replace the context rule, legacy bridge, denominator, or sample without rebuilding the other parts.

Its main limit is conceptual: the data measure whether a filer names a standard, not whether the standard applies. A company can apply a rule without naming it. An industry-specific standard can look rare in a market-wide rate, while a universal standard may need no discussion in the notes. The released `asc_deadletter` table provides two checks. It measures use within the industry governed by a topic and identifies companies that tag the topic's XBRL elements without naming it. Of the 90 topics in force throughout 2020–2024, 48 are cited by fewer than 1 percent of filers. After both checks, 31 remain. These checks can show that a topic is in use; they cannot prove that an uncited topic is unused.

The reverse also occurs: a company may name a standard before it applies or after it has been replaced. ASC 606 receives substantive references from 66 percent of filers in fiscal 2018 and 42 percent in 2024, although it applies to all of them. The series measures attention rather than applicability. Transition notes create a similar issue because they name old and new guidance together: ASC 605 and ASC 606 both peak in fiscal 2018 under either citation rate. When a standard has different dates for large and small filers, the substantive series can show both steps. For CECL (ASC 326), those steps appear in 2020 and 2023, with the second less than half the first. The unfiltered series shows only the first because it already counted smaller filers' advance disclosures. First use is not a clean adoption date either. It inherits the classifier's leak of advance disclosures: for ASC 606, more incumbent filers make a first substantive reference in 2014–2017, while the standard was pending, than in 2018, when it applied. First use dates when a filer first names a topic substantively. Substantive citation measures attention. Neither equals the share of companies legally bound by a topic.

The early panel depends on the legacy term list and bridge. Adding surface forms found in the recall review moved several pre-2009 topic rates by several points. Users of pre-2009 data should inspect the bridge vote counts before relying on the level of a single topic.

Sample composition has a larger effect on the modern data than any extraction choice. In fiscal 2025, removing funds, blank-check companies, structured-finance trusts, and filers without SIC codes changes at least 0.5 percentage points for 30 of the 44 topics cited by at least 1 percent of filers. The change is +9.6 points for revenue and +7.1 for leases. The operating sample ships separately for this reason.

Three smaller limits remain. Delinquent filings assign 1.5 percent of ASU mentions to a fiscal year before the update existed. Originals and amendments both count, so 6.5 percent of filer-period cells contain more than one document. This duplication affects mention-weighted figures but not filer-weighted rates. Finally, the XBRL comparison covers fiscal 2024–2026, and no comparable XBRL data exist before 2009.

An element counts toward a topic only when that topic is its sole authoritative reference, which leaves 54 topics. The rule removes all 23 industry topics (ASC 900–999), including ASC 985, Software, because generic elements such as StockholdersEquity refer only to an industry topic’s illustrative statements. Of the 56 other topics with mapped elements, 16 keep fewer than half; ASC 360 keeps 12 of 32. A low tagging rate for these topics can reflect the rule rather than what filers tag.

V. FUTURE RESEARCH OPPORTUNITIES

The released data support several questions beyond the descriptive analysis in this paper.

Testing guidance-length weights within companies. Prior complexity measures weight a company’s GAAP exposure by the length of applicable standards. Section IV shows that guidance length predicts citation rates across topics. The filer-topic-year table permits a stricter test within companies: do companies exposed to more guidance name more standards? After a CIK and fiscal-year join to Compustat, researchers can also test whether the number or mix of named topics predicts restatements, audit fees, analyst forecast errors, or SEC comment letters.

Measuring advance disclosure under SAB 74. A reference to an issued but not yet effective standard often reports its expected effect. Prior work shows that this disclosure can predict adoption outcomes for one standard (Ahn et al. 2021). The ASU table extends the setting to every update since 2009 and records each effective date. Researchers can compare the breadth and timing of advance disclosure across standards, test whether broad anticipation predicts broad effects, and study companies that disclose before their peers.

Studying the 2009 change in names. The Codification required all filers to replace the names of cited standards at the same time. Legacy names then declined from 57.0 percent of citing filers in 2010 to 15.6 percent in 2013, but never disappeared. The `codification_transition` and `asc_legacy_year` tables can identify which companies rewrote their notes, which kept old language, and whether those choices predict other reporting practices.

Comparing naming with XBRL tagging. For years after 2009, researchers can compare the topics a filer names with those implied by its XBRL tags. This paper reports 2024–2026 only. Joining the SEC Financial Statement Data Sets through the taxonomy reference linkbase would produce an annual panel and show when naming and tagging diverge for each topic. The result could refine the use of XBRL element counts as a complexity measure.

Using the released cross-tabulations. The release provides results by SIC industry, state, domicile, exchange, and form without analysing each split. These tables can measure use of industry guidance within the relevant industry, compare the U.S. GAAP vocabulary of foreign and domestic filers, and separate the standards named in annual and quarterly reports.

Applications in standard setting and audit. FASB post-implementation reviews and SEC filing reviews both ask which registrants a standard reached and when. The ASU table gives annual timing, and the cross-tabulations identify industries and forms. XBRL taxonomy designers can also compare naming with tagging to find topics discussed without a monetary element and topics tagged without an explicit reference.

Extending the extraction. The term list and corpus-building code can be adapted to references to SEC regulations, Public Company Accounting Oversight Board (PCAOB) standards, or IFRS in the same filings. Users can also rebuild the legacy bridge from the released transition sentences. A recall test that downloads a fresh filing sample from EDGAR, rather than reading it from the corpus store, would test ingestion and close the remaining gap in the present validation design. The release includes the sampler and detector needed for that test.

VI. CONCLUSION

This paper releases a dataset of the accounting standards named in periodic SEC filings for fiscal years 2000–2026. It links references before and after the 2009 Codification to common topics, labels the context of each mention, and calculates rates with annual filer counts drawn from the same corpus. The release also includes the code, reference tables, construction evidence, and validation results needed to reproduce or revise the measures.

Prior research has inferred exposure to accounting standards from guidance length and XBRL tags because direct exposure was difficult to observe. The new data record a related but distinct fact: which rules companies choose to name. Researchers can now compare direct references with established proxies, study the timing of disclosure around new standards, and test company-level links to reporting outcomes. For accounting information systems research, the release provides both a benchmark for XBRL measures and a documented text-extraction resource. For practice, it provides a population-wide record of when and where registrants discuss each standard.

DISCLOSURES

Funding

This research received no specific grant from a public, commercial, or not-for-profit funding agency.

Competing interests

None to declare.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE MANUSCRIPT PREPARATION PROCESS

The authors used large language models, including Anthropic’s Claude and OpenAI’s Codex, to draft and revise text, write and review code, label validation samples under written rubrics, and conduct simulated editorial reviews. The authors reviewed the resulting text, code, and labels and take full responsibility for the published article.

REFERENCES

- Abu-Khadra, H., and D. Olsen. 2023. Toward Automating Shredding Nonprofit XML Files: The Case of IRS Form 990 Data. *Journal of Information Systems* 37 (1): 169–188. <https://doi.org/10.2308/isis-2022-031>
- Ahn, J., U. Hoitash, and R. Schmardebeck. 2021. The Informativeness of SAB 74 Disclosures for Accounting Standard Adoption Outcomes: Evidence from ASC 606. Working paper, SSRN 3835849. <https://doi.org/10.2139/ssrn.3835849>
- Alexander, R., M. Ettredge, M. Stone, and L. Sun. 2011. Are Mandatory Disclosure Decisions Made Strategically? The Case of SAB 74 Estimates Preceding Adoption of FIN 48. *Research in Accounting Regulation* 23 (2): 160–166. <https://doi.org/10.1016/j.racreg.2011.06.001>
- Binz, O., R. Hills, and M. Kubic. 2023. Did the FASB Codification Reduce the Complexity of Applying U.S. GAAP?. *Journal of Accounting Research* 61 (5): 1479–1530. <https://doi.org/10.1111/1475-679x.12480>
- Bochkay, K., S. V. Brown, A. J. Leone, and J. W. Tucker. 2023. Textual Analysis in Accounting: What’s Next?. *Contemporary Accounting Research* 40 (2): 765–805. <https://doi.org/10.1111/1911-3846.12825>

- Bommarito, M. J., and D. M. Katz. 2010. A Mathematical Approach to the Study of the United States Code. *Physica A: Statistical Mechanics and its Applications* 389 (19): 4195–4200. <https://doi.org/10.1016/j.physa.2010.05.057>
- Bommarito II, M. J. 2026. Closing the GAAP: Copyright, Fair Use, and Public Access to FASB's Accounting Standards Codification. *SSRN Working Paper* 64. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7457318
- Bonrath, A., M. Eulerich, and N. Gehlen. 2025. Automated Retrieval and Categorization of SEC Data Using Python. *Journal of Information Systems* 39 (2): 139–155. <https://doi.org/10.2308/isisys-2023-063>
- Calderon, T. G., and L. Gao. 2023. Innovative and Novel Research Datasets Related to Cybersecurity Risk Disclosures: A Research Note. *Journal of Information Systems* 37 (2): 123–128. <https://doi.org/10.2308/isisys-2022-065>
- Chychyla, R., A. J. Leone, and M. Minutti-Meza. 2019. Complexity of financial reporting standards and accounting expertise. *Journal of Accounting and Economics* 67 (1): 226–253. <https://doi.org/10.1016/j.jacceco.2018.09.005>
- Codesso, M. M., R. Hoitash, and U. Hoitash. 2026. Textual Financial Data Repository for Machine Learning, Artificial Intelligence, and Textual Analyses: Major Sections from 10-K, 10-Q, and Financial Statement Notes Extracted Using Shared Python Code. *Journal of Information Systems* 40 (2): 117–137. <https://doi.org/10.2308/isisys-2024-084>
- Coupette, C., J. Beckedorf, D. Hartung, M. Bommarito, and D. M. Katz. 2021. Measuring Law Over Time: A Network Analytical Framework with an Application to Statutes and Regulations in the United States and Germany. *Frontiers in Physics* 9: 658463. <https://doi.org/10.3389/fphy.2021.658463>
- Demers, E., V. X. Wang, and K. Wu. 2024. Measuring Corporate Human Capital Disclosures: Lexicon, Data, Code, and Research Opportunities. *Journal of Information Systems* 38 (2): 163–186. <https://doi.org/10.2308/isisys-2023-023>
- Donelson, D. C., J. M. McInnis, and R. D. Mergenthaler. 2012. Rules-Based Accounting Standards and Litigation. *The Accounting Review* 87 (4): 1247–1279. <https://doi.org/10.2308/accr-10274>
- Dyer, T., M. Lang, and L. Stice-Lawrence. 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics* 64 (2-3): 221–245. <https://doi.org/10.1016/j.jacceco.2017.07.002>
- Filzen, J. J., and K. Peterson. 2015. Financial Statement Complexity and Meeting Analysts' Expectations. *Contemporary Accounting Research* 32 (4): 1560–1594. <https://doi.org/10.1111/1911-3846.12135>
- George, J. F., M. Gupta, G. Giordano, A. Mills, V. Tennant, and C. Lewis. 2023. Publishing the Data from “The Effects of Communication Mode and Culture on Deception Detection Accuracy”. *Journal of Information Systems* 37 (2): 129–136. <https://doi.org/10.2308/isisys-2022-025>
- Guay, W., D. Samuels, and D. Taylor. 2016. Guiding through the Fog: Financial statement complexity and voluntary disclosure. *Journal of Accounting and Economics* 62 (2-3): 234–269. <https://doi.org/10.1016/j.jacceco.2016.09.001>
- Hoitash, R., and U. Hoitash. 2018. Measuring Accounting Reporting Complexity with XBRL. *The Accounting Review* 93 (1): 259–287. <https://doi.org/10.2308/accr-51762>

- Hoitash, R., and U. Hoitash. 2022. A Measure of Firm Complexity: Data and Code. *Journal of Information Systems* 36 (2): 161–172. <https://doi.org/10.2308/isisys-2021-041>
- Li, F. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45 (2-3): 221–247. <https://doi.org/10.1016/j.jacceco.2008.02.003>
- Loughran, T., and B. McDonald. 2014. Measuring Readability in Financial Disclosures. *The Journal of Finance* 69 (4): 1643–1671. <https://doi.org/10.1111/jofi.12162>
- Loughran, T., and B. McDonald. 2016. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research* 54 (4): 1187–1230. <https://doi.org/10.1111/1475-679x.12123>
- Lu, H.-M., Y.-T. Chien, H.-H. Yen, and Y.-H. Chen. 2026. Utilizing Pre-Trained Language Models and Large Language Models for 10-K Items Segmentation. *Journal of Information Systems* 1–21. <https://doi.org/10.2308/isisys-2025-005>
- Olsen, D. H., and A. Hassan. 2026. Optimizing SEC Comment Letter Analysis with the CLARITY (Comment Letter Automated Retrieval and Intelligent Text Yielding) Process. *Journal of Information Systems* 40 (2): 139–155. <https://doi.org/10.2308/isisys-2024-062>
- Peterson, K. 2012. Accounting complexity, misreporting, and the consequences of misreporting. *Review of Accounting Studies* 17 (1): 72–95. <https://doi.org/10.1007/s11142-011-9164-5>
- Scott, R., A. N. Skinner, and K. Valentine. 2024. The Disclosure Quality Consequences of Copying Standard-Setter Guidance. *Review of Accounting Studies* 29 (1): 770–808. <https://doi.org/10.1007/s11142-022-09728-7>
- Wang, Y.-Y. 2022. Python Code and Illustrative Crisis Management Data from Twitter. *Journal of Information Systems* 36 (3): 211–217. <https://doi.org/10.2308/isisys-2022-011>
- Zhou, M., Y. Li, Z. Qiao, and B. Shi. 2024. Employee Ratings and Reviews Data from Glassdoor. *Journal of Information Systems* 38 (3): 93–105. <https://doi.org/10.2308/isisys-2023-003>

A. DATA DICTIONARY

Appendix A defines the fields in the main `asc_topic_year` table. Each cross-tabulation adds a key for its grouping variable and otherwise uses the same fields. The deposit’s `DATA_DICTIONARY.md` file defines every field in every released table. All years are fiscal period years. Topic identifiers are three-character text values, including any leading zero. All rates are percentages.

Table 7: Columns of `asc_topic_year`

Column	Meaning
<code>year</code>	Fiscal period year.
<code>topic</code>	ASC topic, three-digit string.
<code>topic_name, area</code>	From <code>asc_topics.csv</code> ; carried for convenience.
<code>superseded_by</code>	The successor topic where one exists (605 → 606, 840 → 842, 310 → 326 for credit losses, 225 → 220, 430 → 606); empty otherwise.

Column	Meaning
rank_in_year	Rank by n_filers (the unfiltered count) within the year. Figure 2 of the paper orders on the substantive rate, so a topic's position there and its rank_in_year can differ.
n_mentions	Mentions, de-duplicated by span, after resolution to this topic. Depth.
n_filings	Distinct accessions with at least one mention.
n_filers	Distinct CIKs with at least one mention. Breadth, unfiltered.
filers_in_year	All registrants filing a periodic report for the period year, from the reference denominator.
pct_filers	$100 * n_filers / filers_in_year$.
n_mentions_asc, n_mentions_asu	Mentions that reached this topic through a direct Codification reference, and through an ASU that amends it. Legacy-bridged mentions are the remainder.
n_filers_asc, n_filers_asu, pct_filers_asc, pct_filers_asu	The same split at the filer grain, and as shares of filers_in_year.
n_filers_substantive	Distinct CIKs with at least one mention classified substantive.
pct_filers_substantive	$100 * n_filers_substantive / filers_in_year$. The headline measure. Filer-weighted.
n_ctx_substantive, n_ctx_asu_title, n_ctx_not_yet_adopted	Mentions by context class. They sum to n_mentions. n_ctx_asu_title is an upper bound on the title mechanism (see the paper's validation section: about a third of the class's predictions are titles).
pct_ctx_substantive	$100 * n_ctx_substantive / n_mentions$. Mention-weighted, so not the same quantity as pct_filers_substantive: a topic can be substantive for most filers and boilerplate for most mentions.
mentions_per_citing_filing	$n_mentions / n_filings$. Verbosity.
pct_in_table	Share of mentions inside a table cell.
pct_disclosure_section	Share of mentions whose Codification reference names Section 50, Disclosure (ASC 842-20-50), i.e. a disclosure requirement rather than recognition or measurement; zero for topics cited without a section.
year_comparable	False for period years whose 10-K or 10-Q count is under 60% of the neighbouring years' (the partial 1999 and the open final year).

B. CODE AND REPRODUCTION

The accompanying repository contains the complete pipeline in the order used in Section III. The file docs/DATA.md gives the commands, run times, and storage requirements.

1. scripts/ingest_full.sh builds the corpus store in about 12 hours and 77 GB. It runs the authors' ingestion tool over the EDGAR full-index and daily feeds; any fetcher that keeps each filing's primary

document can replace it. `scripts/build_filers.py` builds the filer catalog (SIC, state, exchange) from the SEC's public bulk submissions data.

2. `scripts/extract_and_build.sh` runs both extraction passes in less than one hour. Its first section contains the citation term list.
3. `scripts/build_legacy_bridge.py` and `scripts/build_asu_effective_corpus.py` build the two voted reference tables from the mention extract.
4. `scripts/build_model.py` resolves and classifies the references, then writes every released table. The option `--sample` operating writes the operating-company tables. This step takes about 20 minutes.
5. `scripts/validate_context.py` and `scripts/score_precision.py` reproduce the classifier and precision results from the released samples, labels and judgments. `scripts/validate_recall.py` and `scripts/check_zero_match_edgar.py` need the corpus store or EDGAR and ship with their seeds.

The Compustat join uses CIK and fiscal period year. Compustat stores CIK in `comp.company`. Its `gvkey` then links to CRSP `permno` through the CRSP/Compustat Merged table `crsp.ccmxpf_linktable`. If the filer-level release has been loaded as `asc_filer_topic_year`, the Wharton Research Data Services (WRDS) SQL is:

```
select f.gvkey, f.datadate, f.fyear, a.topic, a.n_mentions_substantive
from asc_filer_topic_year a
join comp.company c on cast(c.cik as integer) = a.cik
join comp.funda f on f.gvkey = c.gvkey
    and extract(year from f.datadate) = a.period_year
where f.indfmt = 'INDL' and f.datafmt = 'STD'
    and f.popsrc = 'D' and f.consol = 'C';
```

Two details affect this join. First, `period_year` is the calendar year of the period of report and combines annual and quarterly filings. For a company whose fiscal year does not end in December, one period year can overlap two Compustat `fyear` values. The column `fye_month` gives the month of the filer's annual report in that year. The SQL above matches the year of `DATADATE`, which is the period year; a user joining on `fyear` instead should use `period_year - 1` when `fye_month` is January through May, as Compustat does. The column is blank when the filer has no annual report in the year. Second, `comp.company` stores one CIK for each firm. A registrant that changed CIK will appear under each identifier in the SEC data.

The deposit excludes the corpus store and complete mention extract because of their size, not their licence. The released scripts reproduce both from the public SEC feed, and a 2,000-row mention sample documents the extract schema. The repository's `ascfd` package renders the paper's figures and tables from the released CSV files. Each deposit has a version identifier; users should cite the version used in their analysis.

INDEX OF SUPPLEMENTAL MATERIALS

Please read the READ FIRST agreement before using the files. The supplement contains:

- `ISYS-ASC-Citations_Data_and_Code.zip`, which contains the CSV tables and data dictionary; reference tables with vote counts and provenance; the scripts listed in Appendix B; and the validation samples, labels, and rubric.

- `Supplement.pdf`, a longer report on the same release. It gives the full specification curve across context filters, denominators, samples, and citation schemes; sensitivity of topic ranks to each choice; the legacy bridge and its error model; limits of the XBRL taxonomy mapping; the supersession and transition series; and a record of defects found during review and the resulting changes to the release.

The materials are deposited under a CC-BY licence at [doi:10.5281/zenodo.0000000](https://doi.org/10.5281/zenodo.0000000) [**DOI pending: CC-BY deposit**]. The code is available at <https://github.com/mjbommar/asc-filing-data>.