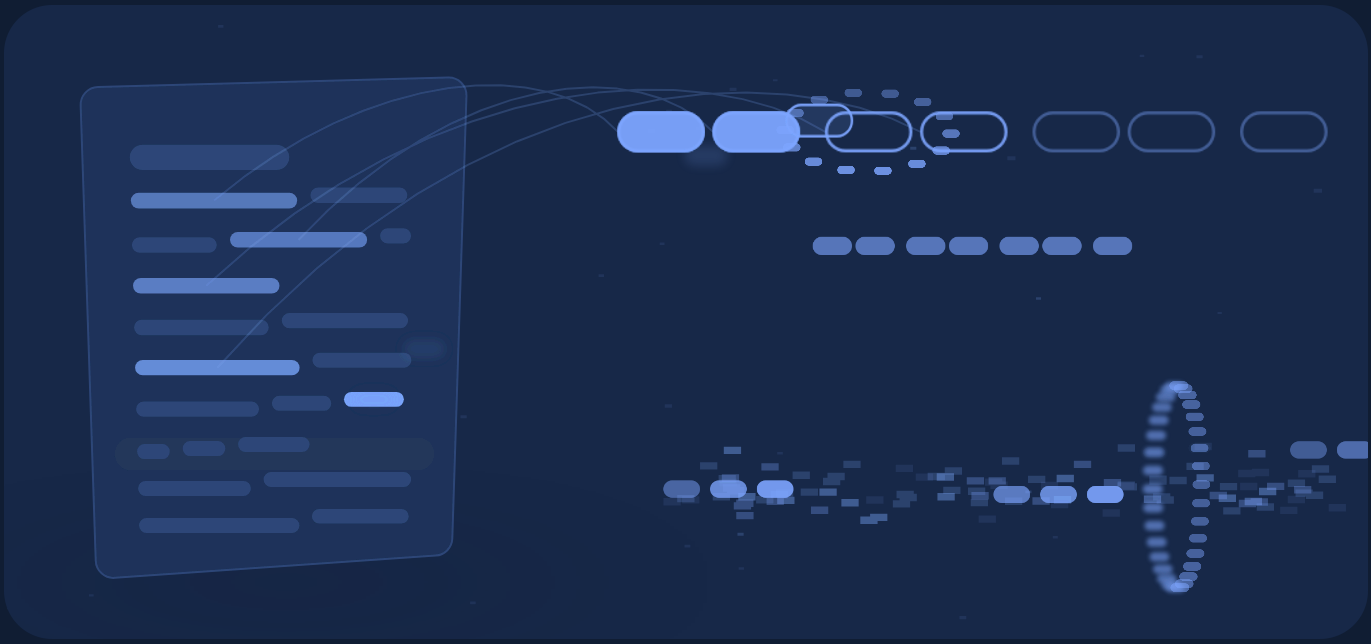
Rillor

EXPLAINER

Extracting data from SEC filings: XBRL and text

Which parts of a 10-K, 10-Q or 8-K must carry XBRL tags, what the data.sec.gov APIs return, and which sections stay as text to extract.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Extracting data from SEC filings starts with knowing which parts the filer had to tag in XBRL and which parts reach EDGAR only as text.

- Regulation S-T Rule 405, the SEC rule on Interactive Data Files, requires Inline XBRL tags on the complete financial statements and footnotes, certain Regulation S-X schedules and a short list of Regulation S-K items, in eCFR text up to date as of 5 November 2026 [1].
- Each footnote, each significant accounting policy and each footnote table is tagged as one block of text, and each amount in a footnote is tagged on its own. Other facts in footnote sentences carry tags only where the filer chooses [1].
- Regulation S-T Rule 405 requires tags on certain Regulation S-K disclosures, including pay versus performance (a five-year table), recovery of erroneously awarded compensation, insider trading policies, cybersecurity and special purpose acquisition company (SPAC) disclosures [1, 10, 17]. Item 106 cybersecurity tags pair a true or false flag with a descriptive text block [23].
- Risk factors, management's discussion and analysis (MD&A), legal proceedings and material contract exhibits are not on Regulation S-T Rule 405's list of tagged items, so their fields have to be read from the filing's text [1, 3, 5, 6, 7].
- The SEC's data.sec.gov XBRL APIs need no key and return only facts that use a standard taxonomy and apply to the whole filing entity [12]. In the companyfacts file of one large accelerated filer, retrieved from those APIs on 8 November 2026, all 25,604 facts were numbers (computed for this report from [12]).
- XBRL values are published as filed, and the SEC does not guarantee its data sets' accuracy [15]. In fiscal 2025 Forms 10-K, custom tags, the elements a filer creates when no standard tag fits, averaged 11.8% of filers' line item tags in the financial statements and 17.0% with the notes [1, 26].
- A material contract exhibit may omit terms the filer customarily and actually treats as private or confidential and that are not material, marked with brackets and a first-page statement [3]. EDGAR full-text search covers filings since 2001, exhibits included [14].
- EDGAR is free, and its fair access rule sets a maximum of 10 requests per second and asks automated users to declare a user agent [13].

Background

What EDGAR holds

EDGAR started in 1994 and 1995, and anyone can access and download its information for free [13]. Its indexes cover all public filings from the third quarter of 1994, by company, by form, as a master list and as a list of submissions with XBRL financial files [13]. The SEC first required XBRL in 2009, at first as a separate XML file and more recently embedded in HTML reports as Inline XBRL [12]. The 2009 release said tagged financial statement information "could be downloaded directly into spreadsheets" [18]. Inline XBRL makes one document both human-readable and machine-readable [17].

How the SEC serves it

The SEC serves filings in four ways: the filing documents with their index files [13], JSON APIs at data.sec.gov [12], the monthly Financial Statement and Notes Data Sets [15] and full-text search over filings since 2001 [14].

A Form 10-K puts risk factors in Item 1A, cybersecurity in Item 1C, legal proceedings in Item 3, MD&A in Item 7, financial statements in Item 8 and exhibits in Item 15 [11]. Whether an item carries tags is set by Regulation S-T Rule 405, which governs the Interactive Data File and leaves the substantive disclosure requirements of the filing unchanged [1].

This report is for analysts, data engineers and compliance teams who want structured data from 10-Ks, 10-Qs and 8-Ks. It describes the structure of public filings and gives no investment advice.

Method

Rule text. Regulation S-T Rules 11, 405 and 406 and Regulation S-K Items 103, 105, 106, 303, 402, 408 and 601 were read through the eCFR versioner API at 5 November 2026, which the eCFR gave on 8 November 2026 as Title 17's "up to date as of" date [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. The list in Regulation S-T Rule 405(b)(4) was compared with the items of Form 10-K [11] and with each Regulation S-K section's own text. An item counts as tagged only where a rule puts its disclosure in an Interactive Data File.

SEC sources. The EDGAR, Inline XBRL, taxonomy and data set pages [12, 13, 14, 15, 16, 17, 24], four adopting releases from 2009 to 2020 [18, 19, 20, 21], the cybersecurity guides [22, 23], analyses and reminders from the SEC's Division of Economic and Risk Analysis (DERA) [26, 27, 28, 29], the 2026 US GAAP taxonomy notes [25], the Inline XBRL specification [30] and one Federal Register notice [35].

Computed checks. On 8 November 2026, with a declared user agent and well under 10 requests per second [13], we queried the companyfacts and submissions files of one large accelerated filer, not named here (in the SEC's documented URL patterns `https://data.sec.gov/api/xbrl/companyfacts/CIK##########.json` and `https://data.sec.gov/submissions/CIK##########.json`), one frames query (`https://data.sec.gov/api/xbrl/frames/us-gaap/AccountsPayableCurrent/USD/CY2025Q4I.json`) [12] and the JSON endpoint behind full-text search (`https://efts.sec.gov/LATEST/search-index`), for exact phrases in documents filed from 1 January to 31 December 2025 [14]. Results are marked "computed for this report".

Research and naming. Two corpus projects [31, 32] and three model studies [33, 34, 36] show how researchers handle filing text. The report names no company, extraction vendor or model.

Findings

1. What Regulation S-T Rule 405 requires in the Interactive Data File

Regulation S-T Rule 11 defines the Interactive Data File as machine-readable code that presents information in XBRL. With Inline XBRL, part of it is embedded in the filing and the rest is submitted as an exhibit [4]. Regulation S-T Rule 405(a)(3) requires Inline XBRL [1]. Its specification, version 1.1 of November 2013 from XBRL International, embeds XBRL fragments in an HTML document that can be viewed in a browser [30].

Content. For an operating company, the file holds a complete set of information for all periods in the filing, "no more and no less", from three categories: the complete financial statements, face and all footnotes, the schedules of Articles 6A and 12 of Regulation S-X, and the Regulation S-K disclosure listed in paragraph (b)(4) [1]. Partial face statements, for example without prior-period comparatives, are not permitted [1].

Accuracy and tags. Each data element, from text and line item names to amounts, dates and labels, must reflect the same information as the filing, and none may be changed, deleted or summarized [1]. Each takes a tag from the standard list in the EDGAR Filer Manual, and a filer creates its own element "if and only if" no appropriate standard tag exists [1]. The SEC's operating company taxonomy page lists 2026 releases of CYD (cybersecurity), ECD (executive compensation disclosure), FFD (filing fees), RXP (resource extraction), SPAC and US GAAP, with SRT and DEI, and its newest IFRS entry is 2025 [24].

Phase-in and forms. XBRL began with periodic reports for fiscal periods ending on or after 15 June 2009, for a first group of filers [18]. Inline XBRL reached large accelerated filers using US GAAP for fiscal periods ending on or after 15 June 2019, accelerated filers from 15 June 2020 and all other operating companies from 15 June 2021 [19]. It covers financial statement and cover page information in Forms 10-Q and 10-K, a Form 8-K's cover page and certain revised financial statements, and certain information in proxy and information statements [17]. Item 601(b)(101) requires a Form 8-K to carry an Interactive Data File exhibit only for revised audited annual financial statements, for the cybersecurity incident disclosure of Item 1.05(b) or for SPAC disclosure under subpart 1600, with a further case for business development companies [3].

2. Block text tagging and detail tagging

Regulation S-T Rule 405(d) sets four levels for footnotes [1].

Level	What is tagged	Required
1	Each complete footnote, as one block of text	Yes
2	Each significant accounting policy in the policies footnote, as one block of text	Yes
3	Each table within each footnote, as one block of text	Yes

Level	What is tagged	Required
4	Each amount within each footnote (monetary value, percentage and number), tagged separately	Yes
4	Each narrative disclosure within a footnote, tagged separately	Only where the filer chooses

Table 1. Footnote tagging under Regulation S-T Rule 405(d), eCFR text up to date as of 5 November 2026 [1].

The 2009 release explained the split. Block text serves users who want an entire footnote or table in context, detail tags on amounts serve users who analyze specific pieces of information, and detail tags on narrative stayed optional [18]. Detailed footnote tagging began in each filer's second year of compliance [27].

In Inline XBRL, the ix:nonNumeric element denotes a non-numeric item, and ix:continuation elements hold further data treated as part of it, so one non-numeric fact can draw on more than one stretch of the HTML [30]. A fact stated in words inside a footnote, such as a condition or a party's name, sits in the block text and needs extraction unless the filer tagged it [1].

The SEC's notes data. The Financial Statement and Notes Data Sets carry "the text and detailed numeric information from all financial statements and their notes", without change from the as-filed reports [15]. They come as eight tab-delimited files: SUB, TAG, DIM, NUM, TXT, REN, PRE and CAL [16]. TXT holds the plain text of non-numeric tagged items, with markup removed and the value truncated to a maximum number of bytes, in a field of 2,048 [16]. Its length field records originals that "may have been greater than 8192" [16]. The documentation reserves this text for text analysis applications [16]. A long footnote needs the filing itself.

3. Regulation S-K items that carry tags

Regulation S-T Rule 405(b)(4) names the Regulation S-K disclosure to be tagged [1], and Items 106, 402 and 408 each repeat the requirement in their own text [8, 9, 10].

Paragraph	Disclosure	Tagging sentence
(b)(4)(i)	Item 402(v), pay versus performance	Item 402(v)(7) [10]
(b)(4)(ii)	Item 402(w), recovery of erroneously awarded compensation, with Form 20-F Item 6.F, Form 40-F General Instruction B(19) and Form N-CSR Item 18	Item 402(w)(4) [10]
(b)(4)(iii)	Item 402(x), awards granted close to the release of material nonpublic information. Item 408(a)(1) and (2), trading arrangements. Item 408(b)(1), insider trading policies. Form 20-F Item 16J(a)	Items 402(x)(3), 408(a)(3) and 408(b)(3) [9, 10]

Paragraph	Disclosure	Tagging sentence
(b)(4)(v)	Item 106 and Form 8-K Item 1.05, cybersecurity, with Form 20-F Item 16K	Item 106 [8]
(b)(4)(vi)	Subpart 1600 of Regulation S-K, SPAC disclosures	Regulation S-T Rule 405(b)(4)(vi) [1]

Table 2. Regulation S-K disclosure that Regulation S-T Rule 405 requires to be tagged, eCFR text up to date as of 5 November 2026. Paragraph (b)(4)(iv) is reserved [1, 8, 9, 10, 17].

The releases behind these items date from August 2022 (pay versus performance) to January 2024 (SPACs) [17].

Cybersecurity. The rules adopted on 26 July 2023 require block text tags on the new narrative disclosures and detail tags on quantitative amounts [22]. All registrants had to begin tagging Form 10-K and 20-F cybersecurity disclosure for fiscal years ending on or after 15 December 2024, and Form 8-K and 6-K incident disclosure by 18 December 2024 [22]. The SEC's March 2026 taxonomy guide shows the form the tags take. A Boolean flag, CybersecurityRiskManagementProcessesIntegratedFlag, records whether risk management processes have been integrated, and a text block holds the description of how [23]. The same pattern runs through the rest of Items 106(b) and (c), and Inline XBRL transformation formats flag a section of text as true or false [23]. For extraction, the flag gives a yes or no and the text block gives the passage. Apart from the true or false flags, which Inline XBRL places on a stretch of the narrative [23], only quantitative amounts are detail tagged, so the description's words sit inside the text block [22].

Where these tags arrive. Inline XBRL also covers certain information in proxy statements [17], and the SEC's taxonomy page lists DEF 14A and PRE 14A submissions [24]. The XBRL APIs list no proxy forms [12], and the companyfacts file checked for this report held no CYD, ECD or proxy statement facts (Finding 5).

4. Cover page tags

Regulation S-T Rule 406, added in 2019, requires filers of Forms 10-K, 10-Q, 8-K, 20-F and 40-F who are required to submit Interactive Data Files in Inline XBRL to tag all the information required on the cover page [2]. Item 601(b)(104) makes the Cover Page Interactive Data File an exhibit [3, 4]. Before the rule, the Form 10-K cover page held about 25 data points, and fewer than half had to be tagged [20]. The untagged points included the exchange on which securities are registered and the state of incorporation [20].

In the companyfacts file checked for this report, dei held two concepts, EntityCommonStockSharesOutstanding and EntityPublicFloat. The same filer's submissions file carried the state of incorporation, addresses, tickers and exchanges as entity fields (computed for this report from [12]). The other cover page tags sit in the filing's own Inline XBRL [2].

5. What the XBRL APIs return and what they leave out

The data.sec.gov APIs return JSON, and they "do not require any authentication or API keys to access" [12]. They cover each filer's submission history and XBRL financial data from Forms 10-Q, 10-K, 8-K, 20-F, 40-F, 6-K and their variants [12]. The XBRL APIs aggregate only facts that use a non-custom taxonomy, such as us-gaap, ifrs-full, dei or srt, and that apply to the entire filing entity [12]. Custom-tagged facts, and facts for part of the entity, fall outside them.

Endpoint	What it holds	Source
submissions	Filing history by 10-digit CIK with leading zeros: at least one year of filings or 1,000 filings, whichever is more	[12]
companyconcept, companyfacts	Facts aggregated from across one filer's submissions	[12]
frames	One fact for each reporting entity, the one "last filed that most closely fits the calendrical period requested"	[12]
companyfacts.zip, submissions.zip	Bulk files, republished nightly at about 3:00 a.m. ET	[12]

Table 3. The data.sec.gov endpoints [12].

Frames periods are CY#### for annual data and CY####Q# for quarterly data, durations of 365 and 91 days within 30 days either way, and CY####Q#I for instants [12]. The JSON updates in real time, typically under a second after processing for submissions and under a minute for XBRL [12].

One filer's facts. One large accelerated filer's companyfacts file held 25,604 facts: us-gaap with 512 concepts, dei with 2 and ffd with 6 (computed for this report from [12]). Every value was a number, and no concept name ended in TextBlock. The facts came from Forms 10-Q (15,304), 10-K (8,966), 8-K (1,118), 10-K/A (210) and 424B2 (6), filed from 22 July 2009 to 6 November 2026, each with accn, end, filed, form, fp, fy and val, plus start and frame where they apply (computed for this report from [12]).

Repeated values. us-gaap:NetIncomeLoss held 341 facts for 124 distinct periods, 87 of them with a frame, and filings reported more than one distinct value for 5 periods (computed for this report from [12]). The companyfacts file keeps the value each filing reported. Frames keeps one per entity, the one last filed [12]. The frames query for AccountsPayableCurrent at CY2025Q4I returned 3,613 data points, one per entity, each with accn, cik, entityName, loc, end and val (computed for this report from [12]).

Data sets beside the APIs. The Financial Statement and Notes Data Sets are monthly since November 2020, quarterly before, and ran from January 2009 to October 2026 when viewed [15]. Their TAG file holds custom tags as well as standard ones [16], which the APIs leave out [12]. The SEC cannot guarantee the data sets' accuracy and calls them no substitute for the filings [15].

6. Items with no tags: risk factors, MD&A, legal proceedings

Form 10-K item	Content set by	Tagged
1A Risk Factors	Item 105 [5, 11]	No [1, 5]
1C Cybersecurity	Item 106 [8, 11]	Yes [1, 8]
3 Legal Proceedings	Item 103 [6, 11]	No [1, 6]
7 MD&A	Item 303 [7, 11]	No [1, 7]
8 Financial Statements and Supplementary Data	The financial statements and footnotes [1, 11]	Yes, footnotes as block text with every amount tagged [1]
15 Exhibits and Financial Statement Schedules	Item 601 [3, 11]	Exhibits 101 and 104 are the XBRL files. Material contracts are not tagged [1, 3]

Table 4. Form 10-K items and their tagging, eCFR text up to date as of 5 November 2026 [1, 3, 5, 6, 7, 8, 11].

Risk factors. Item 105 asks, under the caption "Risk Factors", for the material factors that make an investment speculative or risky [5]. Each goes under a subcaption that adequately describes the risk, generic risks go last under "General Risk Factors", and a section longer than 15 pages gets a summary of no more than two pages [5]. Smaller reporting companies need not provide Item 1A [11]. The subcaptions mark each risk's boundary, and no tag does [1, 5].

MD&A. Item 303 asks for material information to assess financial condition and results of operations, including the amounts and certainty of cash flows [7]. It focuses on known material events and uncertainties reasonably likely to make reported results not necessarily indicative of future results [7].

Legal proceedings. Item 103 names its own record: the name of the court or agency, the date instituted, the principal parties, a description of the factual basis alleged and the relief sought [6]. A damages claim that does not exceed 10 percent of current assets need not be described, and the item may be met by cross-reference to MD&A, risk factors or the notes [6]. A description placed in a footnote falls inside that footnote's text block [1, 6].

How researchers cut the text. EDGAR-CORPUS holds 10-K reports from 1993 to 2020, 6.5 billion tokens from 38,009 companies, with tables removed, HTML stripped and items split by regular expressions [32]. Its authors note that extracting items from documents of hundreds of pages usually requires manual work [32]. The University of Notre Dame's Software Repository for Accounting and Finance finds that a substantial portion of an EDGAR text filing is HTML code, embedded PDFs, images and other artifacts, and some complete filings exceed 400 MB [31]. In one 2012 10-K file, about 7.6% of 48,253,491 characters were the 10-K text, exhibits and tables included [31]. Its parse drops tables, and it found 110 files missing from the fourth quarter of 2015 [31].

7. Exhibits: material contracts and redacted terms

Item 601(b)(10) asks for every material contract not made in the ordinary course of business, for ordinary-course contracts in listed categories, and for management contracts and compensatory plans, which are deemed material when a director or named executive officer takes part [3]. The categories include any contract on which the registrant's business is substantially dependent and any contract for the acquisition or sale of property, plant or equipment for more than 15 percent of consolidated fixed assets [3]. Regulation S-T Rule 405(b)(4) does not list Item 601(b)(10), so contract terms carry no tags [1].

Redaction. A registrant may redact provisions it customarily and actually treats as private or confidential and that are not material [3]. It marks the exhibit index, states on the first page that certain identified information has been excluded, brackets each omission and supplies an unredacted copy on request [3]. Release 33-10618 of March 2019 let registrants make such omissions without a confidential treatment request, for information that was not material and would likely cause competitive harm [20]. Release 33-10884, effective 15 March 2021, replaced the competitive harm test with the customarily and actually treats standard, citing the Supreme Court's decision in Food Marketing Institute v. Argus Leader Media [21].

What full-text search finds.

Exact phrase	Where it comes from	Documents filed in 2025	By form of the parent filing
"certain identified information has been excluded"	Item 601(b)(10)(iv) first-page statement [3]	1,206	8-K 282, 10-Q 203, 10-K 161, S-1 84, DRS 81
"customarily and actually treats"	Standard since 2021 [21]	807	10-Q 180, 10-K 139, 8-K 124
"would likely cause competitive harm"	Standard of 2019 [20]	938	10-K 217, 10-Q 169, 8-K 121

Table 5. EDGAR full-text search, documents filed from 1 January to 31 December 2025. Computed for this report from [14].

For the first phrase, EX-10.1 and EX-10.2 were the top document types (computed for this report from [14]). Each hit is a document, and a phrase can appear outside a legend, so the counts do not count redacted contracts.

8. Data quality: custom tags then and now, and as-filed values

The custom tag rule. Custom elements are allowed only where no appropriate standard tag exists [1], and the SEC has acknowledged that custom tags could potentially reduce the comparability of inter-company data [26].

The trend. DERA's August 2026 analysis covers Form 10-K XBRL under US GAAP for fiscal years 2012 to 2025 [26]. It counts line item tags and excludes abstract, domain member and domain tags [26].

Fiscal year	Financial statements	Financial statements and notes
2012	11.3%	21.4%
2017	9.3%	17.5%
2019	10.4%	17.3%
2022	12.6%	18.6%
2024	11.6%	16.9%
2025	11.8%	17.0%

Table 6. Custom line item tags as a percentage of filers' tags in Form 10-K, average, SEC DERA chart data, rounded [26].

In fiscal 2025, custom tags in the financial statements averaged 14.5% for non-accelerated filers, 14.0% for smaller reporting companies, 9.8% for accelerated filers and 8.9% for large accelerated filers [26]. With notes, the figures were 18.2%, 17.7%, 16.3% and 15.7% [26].

The 2014 history. DERA's July 2014 observations covered XBRL exhibits from 2009 through October 2013 [27]. In DERA's sample, smaller filers accounted for 96% of filers with custom tag rates above 50% [27]. Of the smaller filers with high rates, 64% were served by the same three third-party providers, and one provider accounted for 33% of all filers with a high rate, which suggested the rate may be "an artifact of the reporting tool or service used" [27].

Staff reminders. DERA lists 28 data quality reminders from June 2016 to September 2026, 4 of them dated 2026, on subjects from scaling and negative values to public float errors and Schedule III (counted for this report from [28]). In Forms 10-K for fiscal years ending in 2024, one filer reported a public float of $78 million in its HTML filing and $78 in its XBRL data [29].

Taxonomy change and as-filed data. The 2026 US GAAP taxonomy added 120 elements and deprecated 564, which stay in two annual updates and should not be used in filings [25]. Its notes say the DQC Rules Taxonomy published with it holds 90 of the XBRL US Data Quality Committee's rules, 30 of them added for 2026 [25]. The data sets documentation calls its input uncorrected, as-filed submissions that "may contain redundancies, inconsistencies, and discrepancies" [16].

9. Access rules: request rates, user agent and full-text search coverage

EDGAR information is free, the current maximum request rate is 10 requests per second, and automated users declare a user agent, as in the SEC's sample of a company name and an administrative contact email [13]. data.sec.gov does not support Cross Origin Resource Scripting (CORS), and the SEC gives no technical support for scripted downloading [12]. Full-text search covers electronic filings since 2001 with all attachments, supports Boolean operators, exact phrases, exclusions and wildcards, and does not support natural language search [14]. Filings indexed from 1994 to 2000 fall outside it [13, 14].

Implications

For data engineers. Take numbers from companyfacts and frames with each fact's accession number, form, filing date and period [12]. Choose between values that different filings report for one period by a stated rule, or use frames, which keeps the one last filed [12]. Custom-tagged items are in the Financial Statement and Notes Data Sets and absent from the APIs [12, 16]. Read footnote text from the filing, since the data set's text field is truncated [16].

For analysts. A tagged value is the filer's value as filed [15]. Check it against its passage in the HTML, as the public float case shows [29]. Read custom tag rates by filer group, since they differ: 14.0% for smaller reporting companies and 8.9% for large accelerated filers in fiscal 2025 [26].

For compliance teams. The Regulation S-K items that Regulation S-T Rule 405(b)(4) names for tagging grew with the disclosure rules of 2022 to 2024 [1, 17]. Read the list with the eCFR's "up to date as of" date.

Turning untagged text into source-linked records. The rules give the untagged sections a structure to extract into. Item 103 names the fields of a legal proceedings record [6]. Item 105 gives one risk per subcaption [5]. Item 601(b)(10)(iv)'s brackets mark terms the filer omitted, so a record can show a term as redacted, apart from one the extraction missed [3]. Each field kept with its passage can be checked against the filing by a person.

Model studies support that check. FiNER-139 frames XBRL tagging as entity extraction over 1.1 million sentences and 139 entity types, where the correct tag for a number depends mostly on context [33]. FinTagging finds that LLMs in zero-shot settings generalize well in extraction and struggle significantly with fine-grained linking to more than 10,000 US GAAP concepts [34]. In FinanceBench, a 2023 benchmark of 10,231 questions about publicly traded companies, one model configuration used with a retrieval system incorrectly answered or refused to answer 81% of questions in a 150-case sample reviewed by hand [36].

Where Rillor fits. Rillor reads contracts and filings into one record each, with parties, defined terms, obligations, conditions and deadlines. Each field links to its passage and page, including on scanned copies read with OCR.

Limits of this analysis

- The API checks used one large accelerated filer, and its all-numeric file describes that file only. The API page states only the standard taxonomy and whole-entity conditions [12].
- The phrase counts are documents filed in 2025 that contain a phrase, and they do not count redacted contracts [14].
- The rule text is the eCFR's as of 5 November 2026, and Regulation S-T Rule 405(b)(4) changes with new disclosure rules [1, 17]. The eCFR's head note on the rule links a March 2024 amendment that was delayed indefinitely on 12 April 2024 [1]. That notice delayed the effective date of the climate-related disclosure rules indefinitely, to await judicial review in consolidated proceedings in the Eighth Circuit [35]. The report has no source on those rules after April 2024.
- No source reviewed measures extraction accuracy for risk factors, MD&A, legal proceedings or exhibit terms. The model studies cited cover XBRL tagging and question answering [33, 34, 36].
- The custom tag figures are rounded DERA averages for Forms 10-K under US GAAP [26]. The 2014 figures describe 2009 to 2013 [27].
- The report describes public filings and their rules. It gives no investment or legal advice.

Sources

1. 17 CFR 232.405, Interactive Data File submissions (Regulation S-T Rule 405). eCFR, Office of the Federal Register and Government Publishing Office, text up to date as of 5 November 2026, viewed 8 November 2026. https://www.ecfr.gov/current/title-17/section-232.405

2. 17 CFR 232.406, Cover Page XBRL Data Tagging (Regulation S-T Rule 406). eCFR, Office of the Federal Register and Government Publishing Office, text up to date as of 5 November 2026, viewed 8 November 2026. https://www.ecfr.gov/current/title-17/section-232.406

3. 17 CFR 229.601, Exhibits (Regulation S-K Item 601). eCFR, Office of the Federal Register and Government Publishing Office, text up to date as of 5 November 2026, viewed 8 November 2026. https://www.ecfr.gov/current/title-17/section-229.601

4. 17 CFR 232.11, Definition of terms used in Regulation S-T. eCFR, Office of the Federal Register and Government Publishing Office, text up to date as of 5 November 2026, viewed 8 November 2026. https://www.ecfr.gov/current/title-17/section-232.11

5. 17 CFR 229.105, Risk factors (Regulation S-K Item 105). eCFR, Office of the Federal Register and Government Publishing Office, text up to date as of 5 November 2026, viewed 8 November 2026. https://www.ecfr.gov/current/title-17/section-229.105

6. 17 CFR 229.103, Legal proceedings (Regulation S-K Item 103). eCFR, Office of the Federal Register and Government Publishing Office, text up to date as of 5 November 2026, viewed 8 November 2026. https://www.ecfr.gov/current/title-17/section-229.103

7. 17 CFR 229.303, Management's discussion and analysis (Regulation S-K Item 303). eCFR, Office of the Federal Register and Government Publishing Office, text up to date as of 5 November 2026, viewed 8 November 2026. https://www.ecfr.gov/current/title-17/section-229.303

8. 17 CFR 229.106, Cybersecurity (Regulation S-K Item 106). eCFR, Office of the Federal Register and Government Publishing Office, text up to date as of 5 November 2026, viewed 8 November 2026. https://www.ecfr.gov/current/title-17/section-229.106

9. 17 CFR 229.408, Insider trading arrangements and policies (Regulation S-K Item 408). eCFR, Office of the Federal Register and Government Publishing Office, text up to date as of 5 November 2026, viewed 8 November 2026. https://www.ecfr.gov/current/title-17/section-229.408

10. 17 CFR 229.402, Executive compensation (Regulation S-K Item 402). eCFR, Office of the Federal Register and Government Publishing Office, text up to date as of 5 November 2026, viewed 8 November 2026. https://www.ecfr.gov/current/title-17/section-229.402

11. Form 10-K, General Instructions (SEC 1673 (02-25)). U.S. Securities and Exchange Commission, modified 10 September 2026, viewed 8 November 2026. https://www.sec.gov/files/form10-k.pdf

12. EDGAR Application Programming Interfaces (APIs). U.S. Securities and Exchange Commission, 6 June 2024, last reviewed or updated 8 April 2025, viewed 8 November 2026. https://www.sec.gov/search-filings/edgar-application-programming-interfaces

13. Accessing EDGAR Data. U.S. Securities and Exchange Commission, 23 March 2021, last reviewed or updated 26 June 2024, viewed 8 November 2026. https://www.sec.gov/search-filings/edgar-search-assistance/accessing-edgar-data

14. EDGAR Full-Text Search FAQ. U.S. Securities and Exchange Commission, undated, viewed 8 November 2026. https://www.sec.gov/edgar/search/efts-faq.html (phrase query used for this report: https://efts.sec.gov/LATEST/search-index?q=%22certain%20identified%20information%20has%20been%20excluded%22&dateRange=custom&startdt=2025-01-01&enddt=2025-12-31)

15. Financial Statement and Notes Data Sets. U.S. Securities and Exchange Commission, data January 2009 to October 2026, last reviewed or updated 31 October 2026, viewed 8 November 2026. https://www.sec.gov/data-research/sec-markets-data/financial-statement-notes-data-sets

16. Financial Statement and Notes Data Sets (documentation). U.S. Securities and Exchange Commission, 26 August 2017, viewed 8 November 2026. https://www.sec.gov/files/aqfsn_1.pdf

17. Inline XBRL. U.S. Securities and Exchange Commission, 14 June 2016, last reviewed or updated 17 January 2025, viewed 8 November 2026. https://www.sec.gov/data-research/structured-data/inline-xbrl

18. Release No. 33-9002, Interactive Data to Improve Financial Reporting. U.S. Securities and Exchange Commission, 30 January 2009, viewed 8 November 2026. https://www.sec.gov/files/rules/final/2009/33-9002.pdf

19. Release No. 33-10514, Inline XBRL Filing of Tagged Data. U.S. Securities and Exchange Commission, 28 June 2018, viewed 8 November 2026. https://www.sec.gov/files/rules/final/2018/33-10514.pdf

20. Release No. 33-10618, FAST Act Modernization and Simplification of Regulation S-K. U.S. Securities and Exchange Commission, 20 March 2019, viewed 8 November 2026. https://www.sec.gov/files/rules/final/2019/33-10618.pdf

21. Release No. 33-10884, Facilitating Capital Formation and Expanding Investment Opportunities by Improving Access to Capital in Private Markets. U.S. Securities and Exchange Commission, 2 November 2020, viewed 8 November 2026. https://www.sec.gov/files/rules/final/2020/33-10884.pdf

22. Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure: A Small Entity Compliance Guide. U.S. Securities and Exchange Commission, 30 August 2023, last reviewed or updated 10 September 2024, viewed 8 November 2026. https://www.sec.gov/resources-small-businesses/small-business-compliance-guides/cybersecurity-risk-management-strategy-governance-incident-disclosure

23. Cybersecurity Disclosure (CYD) Taxonomy Guide. U.S. Securities and Exchange Commission, March 2026, viewed 8 November 2026. https://xbrl.sec.gov/cyd/2026/cyd-taxonomy-guide-2026-03-16.pdf

24. Standard Taxonomies: Operating Companies. U.S. Securities and Exchange Commission, 17 October 2023, last reviewed or updated 16 March 2026, viewed 8 November 2026. https://www.sec.gov/data-research/structured-data/taxonomies-

schemas/standard-taxonomies/operating-companies

25. 2026 FASB GAAP Taxonomies Release Notes. Financial Accounting Standards Board, December 2025, viewed 8 November 2026. https://xbrl.fasb.org/resources/annualrelease/2026/GAAP_Financial_Reporting_Taxonomy_Release_Notes.pdf

26. U.S. GAAP XBRL Custom Tags Trend, with its chart data. SEC Division of Economic and Risk Analysis, 10 August 2026, last reviewed or updated 11 August 2026, viewed 8 November 2026. https://www.sec.gov/data-research/structured-data/us-gaap-xbrl-custom-tags-trend (chart data: https://www.sec.gov/files/2026-08/US_GAAP_XBRL_Custom_Tags_Trend_08-04-2026.json)

27. Staff Observations of Custom Tag Rates. SEC Division of Economic and Risk Analysis, 7 July 2014, viewed 8 November 2026. https://www.sec.gov/dera/reportspubs/assessment-custom-tag-rates-xbrl.html

28. Office of Data Standards and Innovation Staff Observations, Guidance, and Trends. SEC Division of Economic and Risk Analysis, 30 May 2024, last reviewed or updated 9 September 2026, viewed 8 November 2026. https://www.sec.gov/data-research/structured-data/office-data-standards-Innovation-staff-observations-guidance-trends

29. Public Float Tagging Errors. SEC Division of Economic and Risk Analysis, 21 July 2025, last reviewed or updated 21 January 2026, viewed 8 November 2026. https://www.sec.gov/data-research/structured-data/2507-dqreminder-public-float-tagging-errors

30. Inline XBRL Part 1: Specification 1.1. XBRL International, Recommendation, 18 November 2013, viewed 8 November 2026. https://www.xbrl.org/specification/inlinexbrl-part1/rec-2013-11-18/inlinexbrl-part1-rec-2013-11-18.html

31. 10X Stage One Parsing Documentation. Bill McDonald, Software Repository for Accounting and Finance, University of Notre Dame, updated January 2022, viewed 8 November 2026. https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/10x-stage-one-parsing-documentation/

32. EDGAR-CORPUS: Billions of Tokens Make The World Go Round. Loukas, Fergadiotis, Androutsopoulos and Malakasiotis, ECONLP 2021 at EMNLP 2021, arXiv:2109.14394, version 2, 1 October 2021 (first version 29 September 2021), viewed 8 November 2026. https://arxiv.org/abs/2109.14394

33. FiNER: Financial Numeric Entity Recognition for XBRL Tagging. Loukas, Fergadiotis, Chalkidis, Spyropoulou, Malakasiotis, Androutsopoulos and Paliouras, ACL 2022, arXiv:2203.06482, version 2, 19 April 2022 (first version 12 March 2022), viewed 8 November 2026. https://arxiv.org/abs/2203.06482

34. FinTagging: Benchmarking LLMs for Extracting and Structuring Financial Information. Wang, Qian, Peng and others, arXiv:2505.20650, version 5, 17 May 2026 (first version 27 May 2025), viewed 8 November 2026. https://arxiv.org/abs/2505.20650

35. The Enhancement and Standardization of Climate-Related Disclosures for Investors; Delay of Effective Date. Securities and Exchange Commission, Federal Register document 2024-07648, 89 FR 25804, 12 April 2024, viewed 8 November 2026. https://www.federalregister.gov/documents/2024/04/12/2024-07648/the-enhancement-and-standardization-of-climate-related-disclosures-for-investors-delay-of-effective

36. FinanceBench: A New Benchmark for Financial Question Answering. Islam, Kannappan, Kiela, Qian, Scherrer and Vidgen, arXiv:2311.11944, 20 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2311.11944

About this report

Published by Rillor on 9 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Extracting data from SEC filings: XBRL and text. 9 November 2026. https://rillor.com/insights/xbrl-and-text-in-sec-filings

Notices

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Send a sample of the rulebooks, regulations, contracts or filings you work from and the fields you need from them. Contact us at rillor.com/contact?type=document-intelligence or contact@rillor.com. More on document and regulation intelligence: rillor.com/si/document-intelligence