

2022年11月のChatGPT無償公開以降、AIは社会・経済に急速な変革をもたらしている。投資家や調査機関においても、分析や情報収集のプロセスにAIを取り入れる動きが広がりつつある。一方で、統合報告書は本来、人間による読み込みを前提に設計されており、機械可読性への配慮は十分ではないと言われてきた。そこで本調査では、2024年1月から12月末の期間に狭義の統合報告書（※）を発行した日経225構成銘柄210社のうち、PDF形式で入手可能な208社を対象に、AIによる利用可能性を左右する「機械可読性（Machine Readability）」の実態を検証した。

（※）「狭義の統合報告書」とは、IIRC フレームワークなどの統合報告ガイダンスを参考にして制作されている報告書、または冊子やWEB サイトでレポート名を統合報告書・統合レポート等と題されている報告書を指す。

調査 サマリー

- コピー可能なPDF設定としている統合報告書が84%
- ファイルサイズは平均14.6MBで、5～19MBが82%
- 機械によるデータ識別に有効なメタデータをなんとか登録しているものは40%
- 機械が文書を理解するのに有用な機能であるタグ付けを実施しているのはまだ少数派

1. 開示をめぐる機械可読性

制度開示においては、各証券取引所の取り決めのもと、XBRL*1や機械可読性のみならず人間可読性も両立したインラインXBRL（iXBRL）*2やXML*3が主流である。書類の要素をタグ*4付けすることでデータの構造がわかるようにし、開示を標準化することに主眼が置かれているが、データのタグ付け作業が企業にとってコスト負担となることが認識されている。*5

機関投資家においては、State Street社の2023年10月の調査*6によると、大半の回答者はまだデータ管理や利用の向上（日本で言うDX）は途上であるが、効率的にデータを扱えるようにDXを進めることが事業機会の拡大に大きく影響すると思うと答えており、2025年の現在においては、機関投資家のDXはだいぶ進んでいるのではと推測される。

一方、任意開示である統合報告書の大多数はPDFで開示されており、PDFは人間可読性を重視したフォーマットであり、機械可読性や転用可能性は低いとされる。また、行政や国際的なプラットフォームに根差したXBRLのようなフォー

マットと違い、PDFはISO等の規格が制定されたものの、私企業であるAdobe社に帰属しており、その管理や技術的主導権が特定の企業に依存しているというガバナンス体制の面で前提条件が異なる。*7

さらに、機械可読性が向上すると、人間可読性に影響が出るのではないかと、個人投資家は恩恵を受けないのでは、という問題提起も見られる。*8

2. 機械可読性やPDFに関連する基準

PDFは人間可読性を追求したフォーマットである。そして統合報告書は、ナラティブ（説明的記述）*9、つまりストーリー性を重視した媒体として、人間が目視することを想定した開示であり、日本ではほとんどがPDFで開示されている。しかし、機関投資家がデータ利活用のグレードアップに取り組む中、統合報告書も、機械可読性を考慮したほうがよいのではないかと考えた。そこで、PDFにおける機械可読性の概念をひもとくため、機械可読性やPDFをめぐる世界の基準について調べてみた。特定した基準は表1の通り。

表 1 PDF及び機械可読性に関連する主な基準

基準名	ISO 32000 (PDF の国際標準規格)	PDF/UA (PDFユ ニバーサル・アクセ シビリティ)旧ISO 14289-1) 国際標準規格	EU「オープンデー タ指令 (Open Data Directive (2019/1024)) 」	US Financial Data Transparency Act of 2022	デジタル庁「オー プンデータ基本指 針」(日本)
URL	https://opensource.adobe.com/dc-acrobat-sdk-docs/pdfstandards/PDF32000_2008.pdf	https://pdfa.org/resource/iso-14289-pdfua/	https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L1024	https://www.congress.gov/bills/117/congress/senate/bills/4295/text	https://www.digital.go.jp/assets/contents/node/basic_page/field_ref_resources/f75de1d-f7ca-4b2a-9b25-94b8a701a037/f1e42cee/20240705_resources_data_p_useline_01.pdf
制定年度 (改訂)	PDF Ver1.7 2008、 (PDF Ver2.0 2020)	2012 (2014, 2024)	2019	2022	2024
基準の目的	PDFの使用環境を 問わない互換性、 視認性、保管可能 性の確保	PDFのユニバーサ ル・アクセシビリ ティ向上	官民データの利活 用の向上	法定開示書類にお ける電子検索性、 AI利用の向上	官民データの利用 の容易化
機械可読性 の言及	×	×	○	○	○
PDFの言及	○	○	×	×	×
AI言及	×	×	○	○	×
メタデータ *10言及	○	○	○	○	○
タグ言及	○	○	×	×	×

AI登場前の時代においては、PDF品質確保のための規格としての観点や、障害者を含め誰もが利用しやすいアクセシビリティに寄与するフォーマットの観点を基準が主流であったと見受けられる。2022年のChatGPTの無償公開前後から、データ利活用（＝機械可読性）を国家競争戦略の文脈でとらえた基準が策定され、日本でも2024年の日本のデジタル庁「オープンデータ基本指針」策定に至っている。

「機械可読性」を主眼としたXBRL等に対し、「PDF」が人間可読性を重視しており、両者は狙いが異なるフォーマットであるためか、「機械可読性」と「PDF」両方に言及した基準は見当たらない結果となった。しかし、アクセシビリティが主眼の基準における利用しやすいPDFの定義は、例えば視覚障害者向けに内容を読み上げてくれるスクリーンリーダーが、書類内の表の構造を理解し、視覚障害者が理解できる形で伝えられるかと

いう観点で定められており、機械可読性に特化したXBRLやXMLが、書類の要素にタグ付けすることでデータの構造がわかるようにする、という仕組みと共通しており、AIによるPDFの機械可読性検証にも準用できると考えられる。従って、本題である統合報告書PDFの機械可読性の調査は、AdobeのPDF上のプロパティやアクセシビリティ検証目的で整備された機能を使用する。

3. 統合報告書PDFの機械可読性調査

2024年1月から2024年12月末の期間で狭義の統合報告書の発行が確認できた日経225銘柄210社のうち、PDFの存在が確認できた208社について、AdobeのPDF上の「文書のプロパティ」と「アクセシビリティをチェック」ツールから代表的な項目について設定を調査した。調査項目は表 2. の通り。

表2 統合報告書PDFの可読性調査項目

	Adobe項目	記載箇所	何がわかるか	機械可読性における示唆
A	セキュリティ→文書の変更 1 許可 2 許可しない	文書のプロパティ	企業が読者に文書の変更を許可しているか	AIが読む場合、変更可にしておく読みやすい可能性がある
B	セキュリティ→内容のコピー 1 許可 2 許可しない		企業が読者に文書内容のコピーを許可しているか	AIが読んで結果をサマライズする際にコピー許可しているほうが処理しやすい可能性がある
C	概要→ファイルサイズ (MB数)		ファイルのデータの重さ	AIが読み込む際に、データが軽いほうが読み込みやすい可能性がある
D	概要→概要 (メタデータ) 0 ファイル名のみ (必須) 1 タイトルあり 2 作成者あり 3 サブタイトルあり 4 キーワードあり		PDFがこういったデータであるか中身を見なくてもわかる	AIが膨大なデータの中から、企業のデータを識別し特定しやすくすることで、読んでもらえる可能性が高まる
E	概要→PDFのバージョン		ISO 32000により初めて標準化されたバージョンであるVer1.7を使用しているか	ISO認証のバージョンとすることで互換性が高くAIにとっても利用しやすい可能性がある
F	画像のみのPDF 1 合格 (テキストあり) 2 不合格 (画像のみ)	アクセシビリティレポート	全体が文字認識できない画像のみとなっているか	AIが文字認識できず内容を読み取れない
G	フォントの埋め込み (埋め込みサブセット) 1 されている 2 されていない	文書のプロパティ	Adobeが認識可能なフォントで埋め込まれているか	埋め込みフォントでない場合、AIが内容を読み取れない
H	優先言語 1 合格 2 不合格	アクセシビリティレポート	書類がどの言語で作成されているか特定されている	AIが読む際に言語が特定されていると処理しやすい可能性がある
I	タグつきPDF 1 合格 2 不合格		タグつきPDFとしての基準を満たしタグつき情報がある	AIが書類の内容を読み取りやすい
J	論理構造 (タグ) → タグ付きコンテンツ (全て) 1 合格 2 不合格		全ての要素にタグが付されており、構造を把握でき、読み上げ順序がわかる	書類の構成を含めてAIによる内容の理解度が高くなる
J'	論理構造 (タグ) → タブの順序 3 合格 4 不合格			
K	代替テキスト→ 図の代替テキスト 1 合格 2 不合格		画像や図にそれがどんな情報であるか説明する文が埋め込まれている	文字がない画像、構造が複雑な図は代替テキストがあることで、何を言いたい図版かAIに伝わりやすい可能性
L	表やグラフの構造→ テーブル→ 行 1 合格 2 不合格		テーブル (表) の要素がタグ付けされておりそれぞれどこに当るかわかるようになっている	文章でなく単語や数値の羅列となる図表では、AIがその構造を理解するには表の構成データがタグにより構造化されていることが必須と思われる
L'	表やグラフの構造→ テーブル→ テーブルヘッダー (TH) とテーブルデータ (TD) 3 合格 4 不合格			
番外編①	目次のリンク 1 あり 2 なし	手動確認	目次が書類内にリンクとなっているか。 (※Adobe上ではリンクの有無に加え、同じリンクが繰り返される場合、スキップする選択肢があるか、リンクが多すぎないかがポイントであるが、検証しやすいようリンク有無とした)	アクセシビリティ上はスクリーンリーダーがアクセスできるURLが重視される (目次から書類内にリンクがあることのAIにとっての利便性は未知数)
番外編②	詳細設定→ アクセシビリティ→ PDF/UA準拠 1 なし 2 UA ISO14289-1準拠	文書のプロパティ	PDF/UA規格に則っているか	ユニバーサル・アクセシビリティの観点なので、AI可読性とどこまで一致するかは不明。(参考まで調査)

4. 調査詳細と考察

(1) セキュリティ

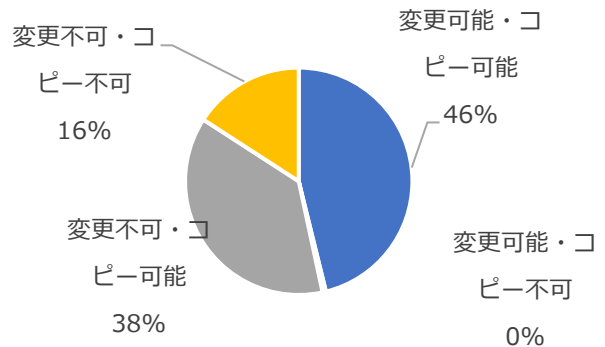
- AIがファイル内を検索するのみならず、情報を拾って分析したレポートに転記するには、PDFファイル上の情報がコピー＆ペースト（コピペ）できるようになっている必要がある。
- 「B 内容のコピー」は174社（84%）が可能としており、狙いがどうであれ、AIが分析するべく転記しやすい形式としている先が大多数となった。
- コピペの可否には影響しないが、「A 文書の変更」は、書類の改変を禁止する指定であり、セキュリティ上重要と考えられ、どのように指定されているか調べたところ、結果は拮抗しており、変更可能97社(47%)と変更不可111社（53%）に分かれた。

考察：情報の正統性維持のため企業のスタンスとしては改変を禁じることは当然と考えられる一方、本当にリスクがある場合、変更不可が大多数となると推測される。機械可読性を念頭に置いた場合、変更可否はどちらのほうが利便性が高いのかは検討の余地があると考えられる。

- 「A 文書の変更」と「B 内容のコピー」の組合せについても分析したところ、「図3 変更・コピー可否」の通り、「変更可能・コピー可能」が最も多く96社（46%）、次に「変更不可・コピー可能」で78社（38%）であり、「変更不可・コピー不可」は33社（16%）であった。

考察：「変更可能・コピー可能」が一番多いことが、セキュリティ意識が低い結果なのか、AI可読性を意識してのことかは、意見が分かれるところである。

図3 変更・コピーが可能なものと不可のものの割合（n=208社）

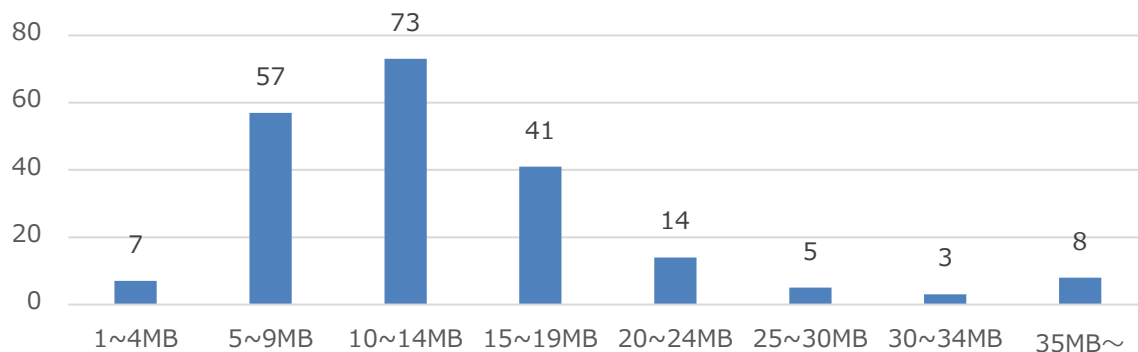


(2) ファイルサイズ

- 考察：**企業のHP上で開いて読むにしても、ダウンロードしてPDFを保存して読むにしても、ファイルサイズが大きすぎると人間でもPCのメモリ容量を消耗し動作環境に影響し利用しづらい。AIが稼働するサーバー上であってもそれは同様と考えられる。
- 「C ファイルサイズ」の結果は、「図4 ファイルサイズの社数分布」の通りであった。平均は14.6MB、最大値は140.7MB、最小値は3.4MB、中央値は68.6MBであった。ボリュームゾーンは5MB から19MB あたりまでと言え、82%を占めた。AIで最大値の140.7MBのPDFを読み込もうとしたところ、そのままでは処理できなかった。

考察：プログラムの仕様次第であろうが、20MB以上のファイルは、処理に時間を要した。人間可読性という統合報告書の主目的に資するビジュアルを犠牲にしない範囲で、ファイルサイズはなるべく軽くすることが望ましいだろう。

図4 ファイルサイズの社数分布 (n=208社)



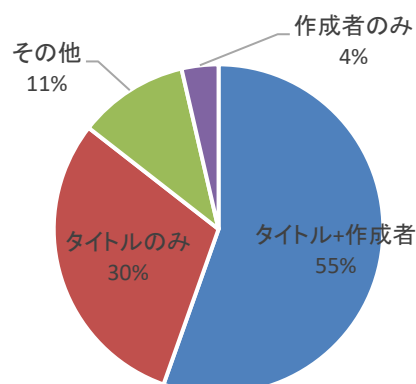
(3) メタデータ*10

- 考察：**人間が統合報告書を読む場合、その企業の統合報告書を読むのだという目的がはっきりしているため、企業はそのPDFのファイル名以外にメタデータとしてタイトル、作成者、サブタイトルやキーワードを登録する必要性はあまり感じないと思われる。しかしPDFの品質目的であれ、アクセシビリティ目的であれ、「表1 PDFにおける機械可読性に関連する基準」のどの基準でもメタデータについて言及されており重要な概念であると考えられる。ましてやAIが無数の開示物の海の中で書類を個別に認識し、分析対象とするか優先順位を付けるとしたらどうであろうか。インターネット検索結果の順位を上げるSEO対策同様、PDFのメタデータを詳細にしておくことは、情報を拾ってもらえる確度に影響するかもしれない。
- 「D メタデータ」**については、なんらか登録していたケースが83社（40%）、登録なしが125社（60%）、と二分されていた。
- 母数を登録ありの83社とした場合のタイトル、作成者、サブタイトル、キーワードの4要素の有無の組合せは「図5 メタデータの組合せの割合」の通りであった。メタデータありの企業83社においては、「タイトル+作成者」の組合せが約半数の46社（55%）、「タイトルのみ」が25社（30%）であった。「タイトル+

作成者」の組合せで登録しているのは、PDFの統合報告書を発行している208社全体の中では22%となった。

考察：AIがデータを識別する上でのメタデータの重要性を考慮すると、「**データ戦略上、意識が高い**」企業群と考えられる。

図5 メタデータの組合せの割合 (n=83社)

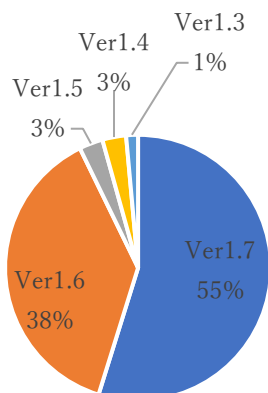


(4) PDFのバージョン

- 考察：**特定のPDFのバージョンで統合報告書を作成しなければならない決まりはないがISO32000により初めて標準化されたバージョンであり互換性が高いとされているため、Ver1.7以降での書類作成が推奨される。
- 考察：**Ver1.7を使用しているのが114社（55%）と半数以上で最多、Ver1.6が79社（38%）であった。多数の企業が順調

にバージョン更新している様子が伺える結果となった。AI可読性においてPDFのバージョンの差異がどのような影響をもたらすかは未知数だが、互換性や利便性を考慮すると、ISOで標準化されたVer1.7に移行するのが得策と考えられる。

図6 PDFのバージョン採用割合 (n=208社)



(5) 文字・言語認識できるか

- PDFが画像だけから成る場合、人間もAIもファイル内の情報を検索することができない。
- 考察：**「F 画像のみのPDF」の結果はゼロ社で、機械可読性を想定せずとも、企業は検索可能性など読者の利便性を考慮していると思われる。
- 「G フォントの埋め込み」は、埋め込み対応していない先は208社のうち1社のみとなった。AIに読ませようとしたところ、当該1社のみ読み込むことができなかった。
考察：AI可読性にとってフォントの埋め込みは必須と考えられる。
- 「H 優先言語」は、指定している先が166社(80%)であり、本来の狙いがどうであれ、AIがどの言語で読めばいいかわかりやすいようにしている先が大多数となった。

(6) テキストの論理構造・図表

- 機械可読性に特化したXBRLやXMLにおいて、データをタグ付けすることで、文章の論理構造やレイアウトがわかるようにすることが前提となっている。AdobeのPDFソフト内のアクセシビリティレポートにおいても、視覚障害者がスクリーンリーダーで読むことを想定し、関連するチェック項目が用意されている。AI可読性においてもこれらは重要であると考えられ、一部項目を調査した。
- 「I タグつきPDF」が合格となっているのは16社(8%)。
- 「J タグ付きコンテンツ(全て)」が合格となっているのは9社(4%)。
- 「J タブの順序」が合格となっているのは3社(1%)。
- 「K 図の代替テキスト」が合格となっているのは7社(3%)。
- 「L テーブル→行」と「L' テーブル→テーブルヘッダー(TH)とテーブルデータ(TD)」が合格となっているのは26社(13%)に上った。
- 文章の論理構造や読み上げ順序、図の代替テキストの埋め込みといったアクセシビリティ上重要な項目は、SECやESMAが指摘するように、企業にとってはタグ付け作業によるコスト負担を意味するためか、1割に満たない項目がほとんどであった。しかし、「L テーブル→行」と「L' テーブル→THとTD」が合格の事例は13%に上った。
考察：テーブル(数表)は財務報告のみならず統合報告書においても重要なコンテンツでありタグ付け慣行が他の制度書類の経験から比較的根付いているのか、それともコスト負担が大きいのかどちらかである可能性がある。AIがテーブルを読む際、役立つことは間違いないと思われるため、これら26社(13%)は、「**データ戦略上、意識が高い**」企業群と考えられる。
- 考察：**「I タグつきPDF」、「J タグ付きコンテンツ(全て)」、「J タブの順序」、「K 図の代替テキスト」が合格の企業は、全て「L テーブル→行」と「L' テーブル→THとTD」も合格となっていたため、テーブルのタグ付けのほうがハードルが低く、対応している「**データ戦略上、意識が高い**」企業群のうち一部が更にコストをかけて他のコンテンツにもタグ付けしているのだと思われる。

(7) 番外編①：目次のリンク

- アクセシビリティ上は、スクリーンリーダーがアクセスできるURLであることに加え、同じリンクが繰り返される場合、スキップする選択肢があるか、リンクが多すぎないかがポイントであるが、検証が複雑であるため、人間可読性にも寄与する観点として、目次にリンクが張っているかを手動で確認した。
- 目次にリンクありは134社（64%）、なしは74社（36%）であり、比較的拮抗した結果となった。

考察：PDFの統合報告書は人間可読性が主眼であるが、枚数が100ページ前後となり、ペーパーレス化の波も考慮すると、目次にリンクがあることは人間の読者の利便性に寄与する。今後更に普及が望まれるであろう。

(8) 番外編②：PDF/UA準拠

- 考察：**PDF/UAはアクセシビリティ向上を目的としたISO規格であり、視覚障害者等のアクセシビリティとAI等による機械可読性は、別の課題であるが、XBRLやXMLにおける論点と、スクリーンリーダー用途における論点に共通のものが多いため。いずれAI可読性のための規格等が整備されるかもしれないが、当座は、PDF/UA準拠等を通して、AI対策とすることも考えられる。
- 「UA ISO14289-1準拠」としている企業は、208社中ゼロであった。

(9) 調査項目間の相関：

- メタデータとセキュリティ
PDFのセキュリティ設定が「変更不可・コピー可能」の割合は全体において78社（38%）であったが、メタデータを「タイトル+作成者」の組合せで登録している「**データ戦略上、意識が高い**」企業群である46社においては、「変更不可・コピー可能」としている企業は、30社（65%）と比率が高くなった。
- テーブルタグ付けとセキュリティ
PDFのセキュリティ設定が「変更不可・コピー可能」の割合は全体において78社（38%）であったが、機械によるテーブルの読みやすさの指標となりうる「L テーブル→行」と「L'

テーブル→THとTD」が合格となっていた「**データ戦略上、意識が高い**」企業群である26社（13%）においては、「変更不可・コピー可能」としている企業は、14社（53%）、「コピー可能」としている企業は23社（88%）と更に比率が高くなった。

- 考察：**メタデータ、テーブルタグ付け、どちらの観点からでも、「**データ戦略上、意識が高い**」企業群の統合報告書は、コピー可能としている割合が高く、転記を含む利用を排除しておらず、AIも分析結果を作成しやすいと考えられる。

9. まとめ

統合報告書は、財務と非財務、過去と将来、社内と社外の情報を統合思考により物語性をもって提示する媒体であり、法定開示のように一律の様式や比較可能性を最優先する書類とは一線を画してきた。そのため、現状の多くは人間の読解に最適化されたPDFであり、AIを含む機械による自動処理を前提とした設計は十分ではない。

現時点で、機関投資家がどの程度AIで統合報告書を読もうとしているか、将来読むか未知数である。AIがどこまで進化するかも未知数である。しかし、調査した企業において、追加コストを要するにも関わらず、タグ付けやメタ情報の活用といった機械可読性の向上に資する工夫を行う企業群も確認された。将来のAIによる企業間比較や要点抽出に有利なデータ戦略をとる、これら「**データ戦略上、意識が高い**」企業群が見られたことは特筆に値する。

機関投資家の多くは、定型的な法定書類を中心にデジタルの開示物は既に機械処理していると思われ、近年のAIの発達を見るにつけ、どんな開示書類であれ、AIが読みやすい形式にしておくことで、読んでもらえる可能性が高まると思われる。ファイルサイズの最適化、PDFバージョン更新、コピー可否設定など低コストで即応できる改善から着手することでまず機関投資家の投資判断の俎上に乗ることが望まれる。加えて、タグ付け・メタデータ整備・構造化といった中長期的な対応を積み重ねることで、人間にとっての読みやすさと機械にとっての読みやすさを並行して高めていくことが、将来を見据えた統合報告書のとるべき形なのではないだろうか。

注記

1. XBRL (eXtensible Business Reporting Language) : 財務報告を含む事業報告の標準化に使用される機械可読性のある構造化データ言語。報告事項にユニークなタグを付すことができ、正確で効率的な報告が可能となる。
2. インラインXBRL (iXBRL) : ウェブページの構造を定義するためのマークアップ言語HTML (HyperText Markup Language)とXBRLタグを使用した構造化データ言語。インターネットブラウザから閲覧可能なことから、専用ソフトを使用せずに機械可読性と人間可読性を両立。
3. XML (Extensible Markup Language) : 任意のデータを補完、伝送、再構築することのできる、タグが埋め込まれたマークアップ言語。異なるシステム間のデータ互換性を重視したもの。
4. タグ : データ上の情報要素に付されるキーワード又は用語で、その情報について説明し、自動的に特定したり処理したりしやすくする機能。
5. U.S. Securities and Exchange Commission, “SEMI-ANNUAL REPORT TO CONGRESS Regarding Public and Internal Use of Machine-Readable Data for Corporate Disclosures” 2025年6月
<https://www.sec.gov/files/2025-fdta-machine-readable-data-report.pdf>
6. State Street, “Capturing the Data Opportunity—Institutional investors in the age of AI” 2023年10月
<https://www.statestreet.com/jp/en/insights/data-opportunity-institutional-investors>
7. European Securities and Markets Authority, “Study on Data Formats for text- based disclosures” 2024年5月
https://www.esma.europa.eu/sites/default/files/2024-07/ESMA12-766636679-320_Study_on_MiCA_Whitepaper_Data_Formats.pdf
8. Andrew C. Call, Ben Wang, Liwei Weng, and Qiang Wu, “Human Readability of Disclosures in a Machine-Readable World” 2024年1月
<https://www.bauer.uh.edu/departments/accy/research/documents/human-readability-of-disclosures-in-a-machine-readable-world-andy-call.pdf>
9. ナラティブ（説明的記述）
IFRS財団「国際統合報告<IR>フレームワーク」2021年1月
4.13組織固有の状況を反映し、論理的である説明的記述
和文 :
https://integratedreporting.ifrs.org/wp-content/uploads/2021/09/IR-Framework-2021_Japanese-translation.pdf
4.13 Narrative flow that is logical given the particular circumstances of the organization
英文 :
https://integratedreporting.ifrs.org/wp-content/uploads/2024/08/IntegratedReporting_Framework_061024.pdf
10. メタデータ : データの特徴を描写する情報。例えば、データフォーマット、ファイルの内容についての説明。

(出所) 株式会社宝印刷D&IR研究所
IR/サステナビリティ研究室の調査による