





# Common Graph Representation of Different XBRL Taxonomies

Artur Basiura<sup>1,2</sup>, Leszek Kotulski<sup>2</sup>, and Dominik Ziemiński<sup>1</sup>

<sup>1</sup> BFT24.COM, ul. Chopina 9/11, 20-026 Lublin, Poland  
{basiura,dominik.ziembinski}@bft24.com

<sup>2</sup> Department of Applied Computer Science, AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Kraków, Poland  
{abasiura,kotulski}@agh.edu.pl

**Abstract.** Information nowadays plays a critical role in our lives, and its misinterpretation or lack of data, makes decisions wrong. It is important to systematize it, not only in the local context but globally. Finance is one of the key areas where standardization and normalization are attempted. One of the attempts is the XBRL format which is widely used in finance. However, the problem is the nature of local implementations. There are many different taxonomies that are implemented independently by countries and organizations. Currently there are no attempts to combine them and create a single standard.

The paper presents a formal model for storing data in graph structures and the concept of using graph grammar to search financial indicators in big data storage. It provides a basis for the future construction of a common graph representation and thus the accumulation of cross-cutting knowledge

**Keywords:** Graph · Graph methods · Xbrl · Similarity graph

## 1 Introduction

Growing global market requires us to make quick and precise decisions. It is not possible without relying on correct and verified data. A perfect example of the misinterpretation of the data is the appearance of the real estate bubble, and consequently the financial crisis of 2007–2008. The reason was not the lack of information but the reliance on incorrect indicators and studies. The scale of the problem touched the globe. In one of the reports on financial crimes [4] it was indicated that reliance on incorrect financial data may be one of the cyber security threats. As a result, access to raw financial data is becoming more and more important. Data that should be verified and available for wider analysis. A popular format that has been implemented in many different countries is XBRL (eXtensible Business Reporting Language) [15]. This format was introduced in 2005 by the SEC (U.S. Securities and Exchange Commission) in order to standardize the reporting structures. Currently, reporting in this format is carried

out not only in the United States but also in Japan, and since the last year in the European Union. The format which in theory unifies the method of settlements at the level of one country does not provide an opportunity to look at the economy globally. It is implemented differently in the United States and in the European Union. In the United States the applicable taxonomy is US GAAP and in the European Union the ESEF. The two taxonomies differ, among other things, with the names of the tags and the way the presentation and clustering layer is organized. The problem is complicated by the fact that it is possible to extend the taxonomy to include own metrics. It is not possible to easily transpose items from one report to another.

This article introduces formal graph notation that is dedicated to the storage of data from the XBRL format. The notation is independent of taxonomy. Importing reports to a graph database gives us the opportunity to perform operations related to the identification of structures with the same similarity [1, 5, 6, 8, 12]. It can be used in practice for very fast data retrieval. The last part shows a practical representation of the report parts in the US GAAP taxonomy. The compilation and storing those data in graph structure allow to search quickly for the needed data. Not only based on one taxonomy.

## 2 XBRL Format and Taxonomies

XBRL (eXtensible Business Reporting Language) is a format used to exchange information between business systems. The language enables the semantic expression of financial and business metrics. This data is stored in an organized manner using XML (Extensible Markup Language) format. Related technologies, such as XML Schema which contains definitions of report components and Namespaces were used to extend the basic structure in order to be able to comprehend the relationships between the report elements and to organize the references. XBRL can be used to store information related to various financial areas or various forms of reporting. To organize reporting issues and provide a single point of reference various taxonomies are defined.

Taxonomy is a system that can be used to identify and structure information that are used to identify financial metrics. It provides information not only about type of structures but also how factors and metrics should be organized (Fig. 1).

To make the XBRL approach consistent, it introduces several ordering layers:

- structure definitions - define what metrics and values may appear in the final report, and what is their meaning. Usually the description comes down to the name of the tag and a short description of the value it represents and the size in which it is expressed (including price tag, currency),
  - description layer - is an extension of the definition with information on how a given tag is to be presented in the context of a specific report, or in the context of the language in which the report is created,
  - presentation layer - that is, the definition of what types of statements may be included in the final report definition, and how they should be presented.
- The main purpose is to be able to present reports in a consistent manner.

```

xhtml:schemaRef xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xbrl="http://www.xbrl.org/2003/linkbase" xlink:type="simple"
xlink:arcrole="http://www.xbrl.org/2003/linkbase" xlink:href="msft-20160630.xsd"/>
<us-gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles contextRef="eol_PE8528----1610-
K0009_STD_0_20141106_0_1333690x1487511" unitRef="iso4217_USD" decimals="-6" id="id_8234187_02F29457-6078-4A0C-89AD-
BF98626D5E09_2002_4">928000000</us-gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles>
<us-gaap:Goodwill contextRef="eol_PE8528----1610-K0009_STD_0_20141106_0_1333690x1487511" unitRef="iso4217_USD" decimals="-8"
id="id_8234187_02F29457-6078-4A0C-89AD-BF98626D5E09_2002_3">1800000000</us-gaap:Goodwill>
<us-gaap:StockRepurchaseProgramAuthorizedAmount1 contextRef="eol_PE8528----1610-K0009_STD_0_20130916_0" unitRef="iso4217_USD"
decimals="INF" id="id_8234187_453ACAEAD-2E70-4B21-A91C-5E6D1D78716_2_0">4000000000</us-
gaap:StockRepurchaseProgramAuthorizedAmount1>
<us-gaap:DebtInstrumentFaceAmount contextRef="eol_PE8528----1610-K0009_STD_0_20151130_0_1324529x1362605" unitRef="iso4217_USD"
decimals="-8" id="id_8234187_F41396E5-A465-40F4-94C2-856A5DC840E8_1001_0">13000000000</us-gaap:DebtInstrumentFaceAmount>
<us-gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles contextRef="eol_PE8528----1610-
K0009_STD_0_20140425_0_1331352x1324667_1333690x1356749" unitRef="iso4217_USD" decimals="-6" id="id_8234187_DD1F40E2-2713-4076-
BAD8-DCFCBF6081D3_2002_1">1500000000</us-
gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles>
<us-gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles contextRef="eol_PE8528----1610-
K0009_STD_0_20140425_0_1331352x1324667_1333690x1356749" unitRef="iso4217_USD" decimals="-6" id="id_8234187_DD1F40E2-2713-4076-
BAD8-DCFCBF6081D3_2001_1">2493000000</us-
gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles>
<us-gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles contextRef="eol_PE8528----1610-
K0009_STD_0_20140425_0_1331352x1324667_1333690x1356749" unitRef="iso4217_USD" decimals="-6" id="id_8234187_DD1F40E2-2713-4076-
BAD8-DCFCBF6081D3_2004_1">157000000</us-
gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles>
<us-gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles contextRef="eol_PE8528----1610-
K0009_STD_0_20140425_0_1331352x1324667_1333690x1356749" unitRef="iso4217_USD" decimals="-6" id="id_8234187_DD1F40E2-2713-4076-
BAD8-DCFCBF6081D3_2003_1">359000000</us-
gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles>
<us-gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles contextRef="eol_PE8528----1610-
K0009_STD_0_20140425_0_1333690x1356749" unitRef="iso4217_USD" decimals="-6" id="id_8234187_9E2C8C5-5765-42CA-87DE-
4B86A6C7FD93_1001_1">147000000</us-gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedIntangibles>
<us-gaap:BusinessCombinationRecognizedIdentifiableAssetsAcquiredAndLiabilitiesAssumedCurrentLiabilities contextRef="eol_PE8528--
--1610-K0009_STD_0_20140425_0_1333690x1356749" unitRef="iso4217_USD" decimals="-6" id="id_8234187_90AD9EE2-3A9D-4893-A058-

```

**Fig. 1.** Example of the fragment of the MICROSOFT report for 2016 in XBRL format [9]

- calculation layer - defines the basic rules for verifying the values that are presented on the report. The layer introduces the calculation rules. Apart from validation they can be used to count the fields that were not included in the report. Usually it has a simple form based on the operation of adding up several indicators.
- formula layer - describes more complex relationships between elements that cannot be described by simple arithmetic operations. Despite the attempt to formalize reports and statements in an orderly manner, it is common to extend the basic definitions introduced by taxonomies by companies that provide reporting. This creates a mess and cannot be easily compared.

All the layers described above are linked to the actual data that is stored in the report file. Single financial data is stored in the form of XML tag in which we have its name, value and additional attributes. For a better understanding of the values and their meaning, a new format has been introduced. InLine XBRL is a format that is based on XBRL tags but allows them to be organized in the form of an unstructured document (HTML/XHTML format) which contain references to XBRL tags. On the one hand these reports are more readable; on the other hand they still contain XBRL fields that can be analyzed by external tools. However, the structure may not be as well defined by the taxonomy. Finding specific information in such data and creating reports can be difficult.

It is also noticeable that each of the above representations can be represented in a graph structure. For example, the calculation rules associated with calculating simple values represent the relationship between tags can take the role of nodes while edges can denote the relationship and the weight with which

elements are associated. The next step is to define the data storage model in a way that is optimized for data processing and retrieval.

### 3 Graph Representation

It is necessary to introduce graph structure in which we store information related to XBRL report data files. It is referred to as a *XBRL graph*.

**Definition 1.** *XBRL Graph* (abbrev. *XG*) is a graph of the form:

$$XG = (V, E, \Sigma, \Gamma, type, attr),$$

where:

- $V$  is a finite, non-empty set of graph nodes,
- $E$  is a finite set of edges,
- $\Sigma$  is a set of node types,
- $\Gamma$  is a set of edge types, where  $\Sigma \cap \Gamma = \emptyset$ ,
- $type : V \cup E \rightarrow \Sigma \cup \Gamma$  is a function that returns the type of a given node/edge:  $type(V) = \Sigma$ ,  $type(E) = \Gamma$ ,
- $attr$  is a function that returns a set of attribute types for a given node/edge type.

We will use the graph to store information from both a single report and a group of reports. Therefore, the following node types have been defined ( $\Sigma$ ):

- *FINT (Financial Indicator Node Type)* - Set of attributes associated with a node are:  $atr(FINT) = \{indicator, value, metric\}$
- *SNT (Statement Node Type)* - The report includes various statements/types of reports (Profit and Loss Account, Balance Sheet and others). The top type determines the type of report that is included in the report. A set of attributes associated with a node:  $atr(SNT) = \{name\}$ .
- *DNT (Document Node Type)* - node type specifying the type of document that represents the report. A set of attributes associated with this type of node:  $atr(DocId) = \{period, report\ type, submission\ date\}$ .
- *CNT (Company Node Type)* - the type of node that identifies reporting institution. Set of attributes associated with a node are:  $atr(CNT) = \{fullname, symbol\}$ .
- *PNT (Period Node Type)* - the type of node specifying the period and date associated with the report or indicator contained in the report
- *TNT (Taxonomy Node Type)* - the type of node that identifies taxonomy which was used to present financial indicators,

This list of node is not complete and can be extended to include specific types depending on the type of analysis to be performed. Proposed graph structure is designed to optimize the search for similar information by year and company type.

We use the following edge types for analysis ( $\Gamma$ ):

- *INCL (Include)* - used to create hierarchical structures. It shows the relationships between the elements of a graph.
- *REL (Related)* - used to model the relationships between elements, the relationships may for example show the calculation rules and how values are related to them.

As with node types, this list is not complete and may be expanded in the future.

Organizing a series of documents using the introduced graph structure gives the possibility to define custom graph grammars and transformations or use mechanisms introduced in other areas. As an example we use graph methods and formal grammars which are used in lighting optimization [1, 7, 10, 11, 13].

### 4 Practical Examples

To better illustrate the concepts, the following is an example of a graph representation of one of the US GAAP taxonomy rules that are represented in XBRL graph form . The income statement is the most common statement in financial statements. One of the representations presented in the US GAAP taxonomy is the calculation rules, which can be written in the form of relationships. The figure visualizes the rules and tag names in an Excel file (shown on Fig 2).

name	label	dept	orde	priorit	weight
GrossProfit	Gross Profit	9	10,0	0	1,0
<b>Revenues</b>	<b>Revenues</b>	<b>10</b>	<b>10,0</b>	<b>0</b>	<b>1,0</b>
SalesRevenueNet	Revenue, Net	11	10,0	0	1,0
FinancialServicesRevenue	Financial Services Revenue	11	20,0	0	1,0
NetInvestmentIncome	Net Investment Income	11	30,0	0	1,0
RealizedInvestmentGainsLosses	Realized Investment Gains (Losses)	11	40,0	0	1,0
RevenuesExcludingInterestAndDiv	Revenues, Excluding Interest and Dividends	11	50,0	0	1,0
InvestmentBankingRevenue	Investment Banking Revenue	11	60,0	0	1,0
UnderwritingIncomeLoss	Underwriting Income (Loss)	11	70,0	0	1,0
MarketDataRevenue	Market Data Revenue	11	80,0	0	1,0
OtherOperatingIncome	Other Operating Income	11	90,0	0	1,0
OtherIncome	Other Income	11	100,0	0	1,0
<b>CostOfRevenue</b>	<b>Cost of Revenue</b>	<b>10</b>	<b>20,0</b>	<b>0</b>	<b>-1,0</b>
CostOfGoodsAndServicesSold	Cost of Goods and Services Sold	11	10,0	0	1,0
FinancialServicesCosts	Financial Services Costs	11	20,0	0	1,0
LiabilityForFuturePolicyBenefitsPer	Liability for Future Policy Benefits, Period Expens	11	30,0	0	1,0
InterestCreditedToPolicyholdersAc	Interest Credited to Policyholders Account Balan	11	40,0	0	1,0
PolicyholderDividends	Policyholder Dividends, Expense	11	50,0	0	1,0
DeferredSalesInducementsAmortiz	Deferred Sales Inducement Cost, Amortization Ex	11	60,0	0	1,0
PresentValueOfFutureInsurancePr	Present Value of Future Insurance Profits, Amorti	11	70,0	0	1,0
AmortizationOfMortgageServicingR	Amortization of Mortgage Servicing Rights (MSRs)	11	80,0	0	1,0
DeferredPolicyAcquisitionCostAmc	Deferred Policy Acquisition Costs, Amortization E	11	90,0	0	1,0
InsuranceTax	Insurance Tax	11	100,0	0	1,0
AmortizationOfValueOfBusinessAc	Amortization of Value of Business Acquired (VOB	11	110,0	0	1,0
OtherCostOfOperatingRevenue	Other Cost of Operating Revenue	11	120,0	0	1,0

Fig. 2. Part of rules for the Revenue tag with calculation base

For example, we can see that statement 124000 defines a GrossProfit indicator, which is dependent on Revenues and CostOfRevenues. Further analysis shows that Revenues depends on a number of different indicators that are extended relative to the underlying taxonomy.

These structures can be represented using XBRL graph, in a hierarchical manner. Thus, all the tags that are in the Microsoft report for the year, are represented as FINT nodes (Financial Indicator Nodes) associated with SNT nodes (Statements), which identify the company and period (shown on Fig 3).

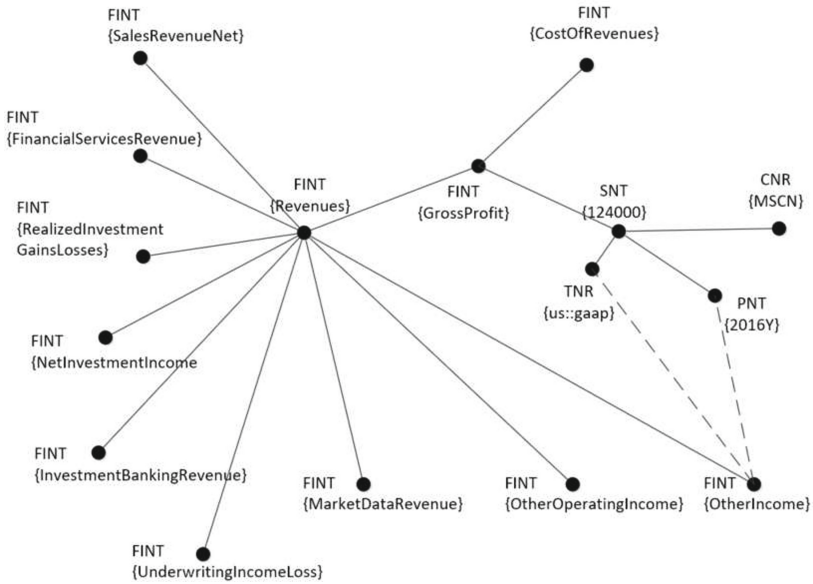
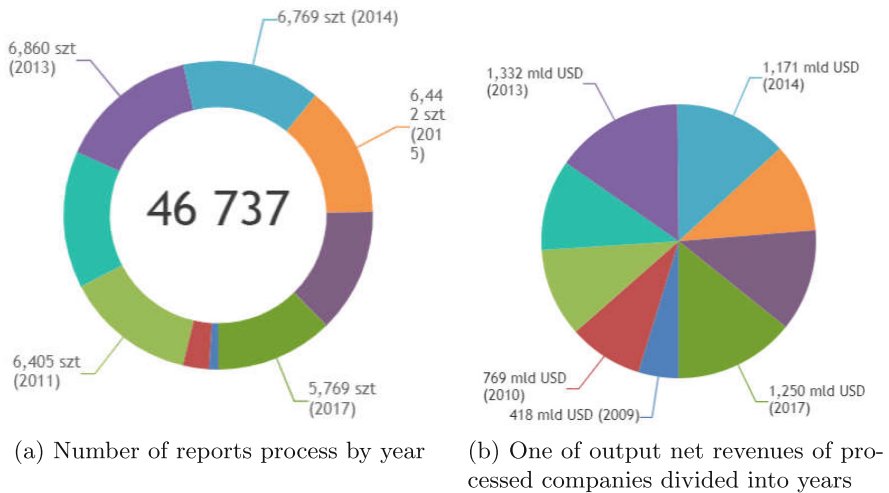


Fig. 3. Part of rules for the Revenue tag with calculation base

On such a graph, it is possible to use graph grammars which are used to divide graphs into smaller graphs to implement agent processing [3, 7, 11, 13]. The created graph can be split to form balanced graphs on which independent functions can be executed.

## 5 Searching XBRL Graph Database

For further analysis, a database was created consisting of reports from 2007 to 2017 containing 46,737 statements, which occupied a space of over 1 Gb of data. From this data, an XBRL graph was created containing 72 million nodes. The current set of reports posted on the SEC website [14] contains over 1 Tb of data. To download data sets, tools provided by BFT24 were used, with which allow to download selected XBRL report files [2].



**Fig. 4.** Results of processing searching Revenues financial indicator

The created database was used to search the set and create simple graph rules, which were aimed at finding the value of revenues and presenting them on a pie chart. To compare the effectiveness, different approaches have been used:

- data processing based on native XBRL sets (XML files has been processed),
- creating XBRL Graph, and then using them to search Net Revenues values,
- creating an XBRL chart, dividing it into 10 subcharts and parallel data search.

The results from the data search were summarized in tabular form, shown in Table 1.

**Table 1.** Time to calculate revenue values per year for the entire dataset

Data storage structure	Time to create a full report
XBRL raw data	557 min
XBRL graph	15.5 min
XBRL graph divided into 10 subgraphs	130 s

There is a very high acceleration in searching for data based on a structured structure, this time can be accelerated when an agent environment is used.

## 6 Conclusions

In this paper, we present a formal model which can be used for storing XBRL reports and case study of processing a group of 46,736 individual XBRL reports.

With the proposed approach it took about 130s to create the report and in comparison with the standard method we reduced the preparation time by more than 10 times. The XG-based method using well-known graph grammars also allows us to perform “what if” analyses. It is possible to analyze datasets and estimate indicators that are not available in reports or to determine groups of errors generated in reports.

Taking into account different types of taxonomies and different methodologies of their creation a common format can also be an element that allows for separation of similar structures and better translation of elements. This will allow to estimate in which set the parameter value will occur.

The concept presented here is also an outline for developing an agent-based system that would offer even faster estimation. The use of parallel processing would allow initial estimates to be obtained in time comparable to real time.

## References

1. Basiura, A., Sędziwy, A., Komnata, K.: Similarity and conformity graphs in lighting optimization and assessment. In: International Conference on Computational Science, pp. 145–157 (2021)
2. BFT24: blockchain financial tools (XBRL processing). <https://prod.bft24.com>
3. Flasiński, M., Kotulski, L.: On the use of graph grammars for the control of a distributed software allocation. *Comput. J.* **25**, 167–175 (1992)
4. Hasham, S., Joshi, S., Mikkelsen, D.: Financial crime and fraud in the age of cybersecurity. <https://mckinsey.com/business-functions/risk-and-resilience/our-insights/financial-crime-and-fraud-in-the-age-of-cybersecurity> (2019). Accessed 06 Apr 2022
5. Komnata, K., Basiura, A., Kotulski, L.: Graph-based street similarity comparing method. In: Zamojski, W., Mazurkiewicz, J., Sugier, J., Walkowiak, T., Kacprzyk, J. (eds.) *DepCoS-RELCOMEX 2020*. AISC, vol. 1173, pp. 366–377. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-48256-5\\_36](https://doi.org/10.1007/978-3-030-48256-5_36)
6. Kotulski, L.: *Rozproszone transformacje grafowe: teoria i zastosowania*. Wyd. Naukowe AGH (2013)
7. Kotulski, L., Sedziwy, A.: GRADIS - the multiagent environment supported by graph transformations. *Simul. Model. Pract. Theory* **18**(10), 1515–1525 (2010)
8. Kotulski, L., Sedziwy, A.: Parallel graph transformations supported by replicated complementary graph. In: ICANNGA, pp. 254–264 (2011)
9. Microsoft: XBRL Raport Data File. <https://sec.gov/Archives/edgar/data/789019/000119312516662209/msft-20160630.xml> (2016). Accessed 06 Apr 2022
10. Sędziwy, A., Kotulski, L., Basiura, A.: Enhancing energy efficiency of adaptive lighting control. In: Nguyen, N.T., Tojo, S., Nguyen, L.M., Trawiński, B. (eds.) *ACIIDS 2017*. LNCS (LNAI), vol. 10192, pp. 487–496. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-54430-4\\_47](https://doi.org/10.1007/978-3-319-54430-4_47)
11. Sędziwy, A., Kotulski, L., Basiura, A.: Multi-agent support for street lighting modernization planning. In: Nguyen, N.T., Gaol, F.L., Hong, T.-P., Trawiński, B. (eds.) *ACIIDS 2019*. LNCS (LNAI), vol. 11431, pp. 442–452. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-14799-0\\_38](https://doi.org/10.1007/978-3-030-14799-0_38)
12. Sedziwy, A.: Effective graph representation supporting multi-agent distributed computing. *Int. J. Innovative Comput. Inf. Control* **10**(1), 101–113 (2014)



13. Sedziwy, A., Basiura, A.: Energy reduction in roadway lighting achieved with novel design approach and leds. LEUKOS: J. Illum. Eng. Soc. North Am. **14**(1), 45–51 (2017). <https://doi.org/10.1080/15502724.2017.133015>
14. U.S. Securities and Exchange Commission(SEC): Electronic Data Gathering, Analysis, and Retrieval system data. <https://www.sec.gov/Archives/edgar/full-index/>. Accessed 06 Apr 2022
15. U.S. Securities and Exchange Commission(SEC): Structured Disclosure at the SEC: History and Rulemaking. <https://www.sec.gov/page/osdhistoryandrulemaking>. Accessed 06 Apr 2022