Einsatzbereiche semantisch strukturierter Daten bei der Unternehmensanalyse

Rechts- und Wirtschaftswissenschaftlichen Fakultät

Fachbereich Wirtschafts- und Sozialwissenschaften

Friedrich-Alexander-Universität Erlangen-Nürnberg

zur

Erlangung des Doktorgrades Dr. rer. pol.

vorgelegt von

Julian Grümmer, M.Sc.

aus Nürnberg

Als Dissertation genehmigt

von der Rechts- und Wirtschaftswissenschaftlichen Fakultät
vom **Fachbereich Wirtschafts- und Sozialwissenschaften**
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung: 27.06.2022

Vorsitzende/r des Promotionsorgans: Prof. Dr. Klaus Henselmann

Gutachter/in: Prof. Dr. Klaus Henselmann

Prof. Dr. Thomas M. Fischer

Vorwort

Die vorliegende kumulative Dissertationsschrift ist während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Lehrstuhl für Rechnungswesen und Prüfungswesen an der Friedrich-Alexander-Universität Erlangen-Nürnberg entstanden. Die Erstellung war für mich Herausforderung und persönlich bereichernde Erfahrung zugleich. Viele Menschen haben mich während der Promotionszeit begleitet und auf unterschiedliche Weise unterstützt. Bei ihnen allen möchte ich mich bedanken. Folgenden Personen gilt jedoch mein besonderer Dank, da sie mich in besonderer Weise bei meinem Dissertationsvorhaben unterstützt haben.

An erster Stelle möchte ich ausdrücklich meinem akademischen Lehrer Herrn Professor Dr. Klaus Henselmann danken, der mich stets gefördert, mir viel Freiraum zur eigenen Entfaltung in Forschung und Lehre schenkte und mich bei meinem Dissertationsvorhaben stets begleitet hat. Ebenfalls bedanke ich mich bei Herrn Professor Dr. Thomas Fischer für die Übernahme des Zweitgutachtens. Der Diskurs und die spannenden Fragen im Rahme der Verteidigung der Dissertation haben mich sehr gefordert. Mein herzlicher Dank gilt auch Professor Dr. Andreas Harth für die gute Zusammenarbeit bei verschiedenen Forschungsprojekten und die vielen fachlichen Anregungen und Diskussionen.

Auch meinen Koautoren möchte ich für die anspruchsvolle und konstruktive Zusammenarbeit danken. Jeder Einzelne hat die gemeinsamen Forschungsprojekte zu einem besonderen Erlebnis gemacht.

Großer Dank gilt auch allen aktuellen und ehemaligen Kolleginnen und Kollegen am Lehrstuhl. Ganz besonders möchte ich in diesem Zusammenhang Dr. Julia Vetter und Dr. Michael Dimmer für die tolle Zusammenarbeit am Lehrstuhl danken - ohne Euch wäre meine Promotionszeit nicht dieselbe gewesen. Ich möchte mich aber auch bei allen studentischen Hilfskräften bedanken, die mich stets unterstützt haben. Danke sagen möchte ich aber auch für die „Ablenkungen“. Diese haben mir auch in anstrengenden Zeiten ein Lächeln auf die Lippen gezaubert.

Lieber Goran, ich danke auch Dir ganz besonders für die sehr lehrreiche Zusammenarbeit an unseren gemeinsamen Forschungsprojekten und freue mich bereits auf alles, was noch kommen mag.

Mein Dank gilt auch allen nicht namentlich genannten Personen. Ihr alle habt mir eine wunderschöne Zeit an der FAU geboten und ich freue mich nicht nur, dass ich Euch kennenlernen durfte, sondern Euch nun auch meine Freunde nennen darf.

Außerdem gilt mein Dank meiner Familie, ohne deren Unterstützung und Rückhalt ich meinen akademischen Weg nicht so unbeschwert und glücklich hätte beschreiten können. Ich möchte mich hier insbesondere bei meinen Eltern bedanken. Danke, dass Ihr mich zu jeder Zeit in außergewöhnlicher und bedingungsloser Art und Weise gefördert habt. Euch sei diese Arbeit gewidmet.

Meine „Reise Promotion“ ist für mich nun beendet und ich stelle fest, dass diese Jahre mich wohl mehr geprägt haben als alle anderen Lebensabschnitte zuvor. Ich blicke mit viel Freude und Stolz auf die letzten Jahre zurück. Nach dieser doch sehr lehrreichen und turbulenten Zeit freue ich mich auf alles, was ich in meinem Leben noch vor mir habe.

Zu guter Letzt gilt mein Dank jedoch Dir. Für Dein großes Verständnis, Deine Zuversicht, Dein Vertrauen und Deine Lebensfreude bin ich Dir unendlich dankbar. Du warst immer für mich da und hast mir stets die nötige Kraft und Unterstützung gegeben, dass ich meine Promotion abschließen konnte. Auch dir widme ich diese Arbeit.

Nürnberg, im August 2022

Julian Grümmer

Geleitwort

Herr Dr. Julian Grümmer untersucht in seiner Arbeit „Einsatzbereiche semantisch strukturierter Daten bei der Unternehmensanalyse“. Ausgangspunkt bildet die Tatsache, dass in Rechnungslegung, Besteuerung, Prüfungswesen und Corporate Governance zunehmend strukturierte Datenformate Verwendung finden.

Während die Formate XBRL/iXBRL auf die Finanzberichterstattung und künftig auch auf Nachhaltigkeitsberichte (ESG Reporting) spezialisiert sind, werden in vielen anderen Anwendungsfeldern strukturierte Daten als sog. Wissensgraphen gespeichert und auch veröffentlicht. Solche Knowledge Graphs sind besonders für dynamisch wachsende Datenbestände von „Objekten“ mit unterschiedlichsten Merkmalen geeignet, die sich schlecht in Tabellen- oder hierarchische Baumstrukturen pressen lassen. Die Daten – etwa über Vorstandsmitglieder, Aufsichtsräte und Abschlussprüfer – sind vielmehr netzwerkartig, also als mathematischer Graph, organisiert.

Vor diesem Hintergrund greift Herr Grümmer in sechs technikaffinen Einzelbeiträgen beide Varianten semantisch strukturierter Daten auf. Er zeigt mögliche Einsatzbereiche von Datenanalysen sowie auch Verbindungsmöglichkeiten zwischen beiden Daten-Welten auf. Dabei bedient er sich forschungsmethodisch primär der Design Science Research (DSR). Diese generiert für gegebene Probleme exemplarische Programmcodes als sog. Artefakte und untersucht deren Beiträge zur Lösung des Problems. Seine vielfältigen Erkenntnisse wurden auf zahlreichen internationalen Konferenzen vorgestellt.

Mit diesen – für ihn typischen – innovativen Themen sind Fragen angesprochen, die auch jenseits des rein wissenschaftlichen Interesses von hoher Relevanz für Gesetzgeber, berichtende Unternehmen, Abschlussprüfer, Betriebsprüfer sowie Abschlussadressaten sind. Seiner Arbeit, die in hervorragender Weise Verbindungen zwischen der Rechnungslegung und der (Wirtschafts-)Informatik herstellt, ist daher eine weite Verbreitung zu wünschen.

Nürnberg, im August 2022

Klaus Henselmann

Abstract

This dissertation consists of 6 papers. The papers investigate the areas of application of semantically structured data in company analysis.

The first paper examines the extent to which alternative data can support tax audits. It is also important to understand how alternative data can be used in the first place. This includes the generation and further the processing of the data. Different methods are presented for this purpose.

The second paper deals with the quality of inline eXtensible Business Reporting Language (iXBRL) annual reports from the UK. Small and medium-sized enterprises (SMEs) have been required to make their annual reports available in iXBRL since 2011. These reports are both machine and human readable. This offers the possibility to evaluate the annual reports of SMEs by machine. The study examines various dimensions of quality such as the structure of the tags. The analysis is important because the use of XBRL and iXBRL will still play an important role in the future, especially regarding the new European Single Electronic Format (ESEF) and the EU taxonomy for sustainable activities.

The third paper analyzes iXBRL company accounts of SMEs in the UK. The information from the annual reports is processed automatically and linked to other data sources with the help of a knowledge graph. In this way, further information can be generated and processed. Other data sources are, for example, another database of the UK Companies House, but also other "alternative data". Linking the data sources using the knowledge graph then enables the data to be queried, evaluated and visualized.

The fourth paper deals with the question how connections between supervisory boards, management boards and auditors can be visualized with the help of the graph database Neo4j. Basically, the problem is that networks are difficult to analyze and display. For this reason, this paper deals with the use of a graph database to close this gap. The database includes supervisory board members, management board members and auditors of DAX30 companies in 2019. With the help of Neo4j, the Curricula Vitae (CVs) are matched and the persons with common activities are analyzed. Among other things, multiple mandates, joint professional activities and joint training are examined.

The fifth paper examines the collapse of the German financial services company Wirecard AG. The focus here is particularly on the people behind the scandal. For this purpose, the CV of the members of the supervisory board and the management board of the DAX30 companies are collected. The information is taken from the respective company websites. In addition, publicly available data sources are used to expand the database. The data is analyzed and visualized with the aid of a knowledge graph. In particular, the special features of the members of the supervisory board and the management board of Wirecard is addressed and the paper examines the differences to other DAX30 companies in detail.

The sixth paper is dedicated to teaching the basics of Resource Description Framework (RDF) and the RDF query language SPARQL. It has already been shown in the previous studies that the use of Knowledge Graphs can bring great benefits. In this paper, the concept of the WireGraph learning game is presented, which is used to teach business students how to work with knowledge graphs in order to work through the Wirecard scandal. For this purpose, a prosumer environment is provided, with which the competence area of digital content creation of the reference model is taught completely and across all performance levels. In doing so, students not only learn the basics of RDF and SPARQL but can also apply them specifically to their ideas and further expertise. The conception of the learning game WireGraph is the first stage of a multistage project. The following steps are the development of the learning game and afterwards the application and the measurement of the didactic success.

Zusammenfassung

Die vorliegende Dissertation besteht aus 6 Beiträgen. Diese Beiträge untersuchen die Einsatzbereiche semantisch strukturierter Daten bei der Unternehmensanalyse.

Die erste Studie untersucht, inwiefern „Alternative Data“ bei der Betriebsprüfung unterstützen können. Hierbei ist es auch wichtig zu verstehen, wie „Alternative Data“ überhaupt genutzt werden können. Dazu zählen das Generieren und die Weiterverarbeitung der Daten. Hierfür werden unterschiedliche Methoden dargestellt.

Der zweite Beitrag beschäftigt sich mit der Qualität von inline eXtensible Business Reporting Language (iXBRL) Geschäftsberichten aus Großbritannien. Kleine und mittlere Unternehmen (KMU) müssen bereits seit 2011 ihre Geschäftsberichte in iXBRL zur Verfügung stellen. Diese Berichte sind sowohl maschinen- als auch menschenlesbar. Die Studie untersucht verschiedene Dimensionen der Qualität, wie beispielsweise die Struktur der Tags. Die Analyse ist wichtig, da die Verwendung von XBRL und iXBRL in der Zukunft, insbesondere im Hinblick auf ESEF und die EU-Verordnung für nachhaltige Aktivitäten, noch eine wichtige Rolle einnehmen wird.

Der dritte Beitrag beschäftigt sich mit iXBRL Geschäftsberichten von KMU in Großbritannien. Die Informationen aus den Geschäftsberichten werden maschinell weiterverarbeitet und mithilfe eines Wissensgraphen mit anderen Datenquellen verknüpft. So können weitere Informationen generiert und verarbeitet werden. Andere Datenquellen sind beispielsweise eine weitere Datenbank des UK Companies House, aber auch andere „Alternative Data“. Die Verknüpfung der Datenquellen mittels Knowledge Graph ermöglicht anschließend die Abfrage, Auswertung und Visualisierung der Daten.

Der vierte Beitrag beschäftigt sich mit der Frage, inwiefern Verbindungen zwischen Aufsichtsräten, Vorständen und Wirtschaftsprüfern mit Hilfe der Graph-Datenbank Neo4j visualisiert werden können. Grundsätzlich besteht das Problem, dass Vernetzungen nur schwer zu analysieren und darzustellen sind. Aus diesem Grund beschäftigt sich dieser Beitrag mit der Verwendung einer Graph-Datenbank, um diese Lücke zu schließen. Die Datenbasis umfasst Aufsichtsräte, Vorstände und Wirtschaftsprüfer der DAX30-Unternehmen im Jahr 2019.

Mithilfe von Neo4j werden die Lebensläufe abgeglichen und die Personen mit gemeinsamen Tätigkeiten analysiert. Hierbei werden u.a. Mehrfachmandate, gemeinsame berufliche Tätigkeiten und gemeinsame Ausbildung untersucht.

Der fünfte Beitrag beschäftigt sich mit dem Zusammenbruch des deutschen Finanzdienstleistungsunternehmens Wirecard AG, insbesondere mit den involvierten Personen. Dafür werden die Lebensläufe der Mitglieder des Vorstands und des Aufsichtsrats der DAX30-Unternehmen gesammelt. Die Informationen stammen aus den jeweiligen Unternehmenswebseiten. Zusätzlich werden öffentlich zugängliche Datenquellen genutzt, um die Datenbasis zu erweitern. Mithilfe eines Knowledge Graphs werden die Daten analysiert und visualisiert. Dabei wird insbesondere auf die Besonderheiten der Mitglieder des Vorstands und des Aufsichtsrats von Wirecard eingegangen und untersucht, ob und welche Unterschiede zu anderen DAX30-Unternehmen bestehen.

Der sechste Beitrag widmet sich dem Lehren der Grundlagen des Resource Description Frameworks (RDF) und der graphenbasierten Abfragesprache SPARQL. Bereits in den vorangehenden Studien konnte gezeigt werden, dass die Verwendung von Knowledge Graphs einen großen Nutzen bringen kann. In diesem Beitrag wird das Konzept des Lernspiels WireGraph vorgestellt, mit dem Studierende der Wirtschaftswissenschaften die Arbeit mit Wissensgraphen erlernen, um den Wirecard-Skandal aufzuarbeiten. Dazu wird eine Prosumentenumgebung gestellt, mit der der Kompetenzbereich *Erstellung digitaler Inhalte* des Referenzmodells vollständig und über alle Leistungsniveaus hinweg vermittelt wird. Dabei erlernen Studierende nicht nur die Grundlagen von RDF und SPARQL, sondern können diese auch gezielt für ihre Ideen und ihr weiteres Fachwissen einsetzen.

Contents in brief

Chapter A

Introduction

Contents – Chapter A

1 Background

This dissertation explores the use of semantically structured data in company analysis. Both the nature of data as well as its use has evolved over time. This has also arrived in financial reporting and company analysis. For example, there is an increasing use of structured data formats. These include, among others, XBRL and IXBRL. The main advantage is that company data can now also be processed by machine (Herren Lee, 2020). XBRL and iXBRL have already been used in financial reporting in several countries for some time. The use of XBRL and iXBRL is not limited to listed companies. In the UK, SMEs are required to file their financial reports in iXBRL format. However, it is important to remain familiar with the reporting format, especially as iXBRL continues to be applied in the EU through the introduction of ESEF and the EU taxonomy for sustainable activities.

Furthermore, there are also new ways how data can be processed. Google has already been using a semantic search engine since 2010 with the start of the Hummingbird ranking algorithm. The term knowledge graph is used to describe a general system for searching and linking information. The basis is a new arrangement of data, which is now no longer hierarchical but network-like, i.e. sorted in the form of graphs (Fensel et al., 2020). Knowledge graphs have become increasingly important in recent years and are also being applied in more and more areas (Gartner, 2020). This is also reflected in the valuation of companies that specialize especially in the processing of data with knowledge graphs, such as the graph database company Neo4j, which recently reached a valuation of $2 billion-plus (Metinko, 2021).

There are two different forms of using knowledge graphs. One is the use of RDF, a W3C standard (RDF, 2014). The other is labeled property graphs, which Neo4j implements (Robinson et al., 2013; Barrasa, 2017). Both are particularly suitable for showing connections between elements as well as for dynamically growing and expanding data sets.

The research method used in the dissertation follows the design science approach which aims to generate new and innovative artifacts that serve to enhance the capabilities of people or organizations (Hevner et al., 2004).

Design science as a design-oriented approach (Becker et al., 2009; Österle et al., 2010) is one of the two central research approaches in business informatics, along with the behavioral approach, which aims to explain or predict the behavior of people or organizations (Wilde & Hess, 2015; Schreiner et al., 2015). In design science, knowledge is generated through the creation of artifacts and an understanding of a problem and its solution is generated (Peffers et al., 2007).

Design Science Research (DSR) can be described as three closely related cycles of activities (Figure 1). This design knowledge helps research and practice to design artifacts systematically and scientifically in future projects. An artifact is a (usually technological) solution to a specific research problem.

The starting point of design science is the relevance cycle which provides the application context (Vom Brocke et al., 2020). The relevance cycle refers to the environment which consists of people, organizations and technology. The relevance cycle specifies the requirements of the research such as the problem to be addressed as input and determines acceptance criteria which help in the evaluation of the research results.

The second cycle which is called rigor cycle provides knowledge from both foundations (e.g., theories, frameworks) and methodologies (e.g., experimentation, data analysis). The appropriate application of both foundations and methodologies achieves a consistent level of rigor (Vom Brocke et al., 2020).

The third and last cycle which is the heart of design science research is the design cycle. The design cycle refers to the production of design alternatives and assessment of the alternatives against the requirements until an appropriate design is achieved.

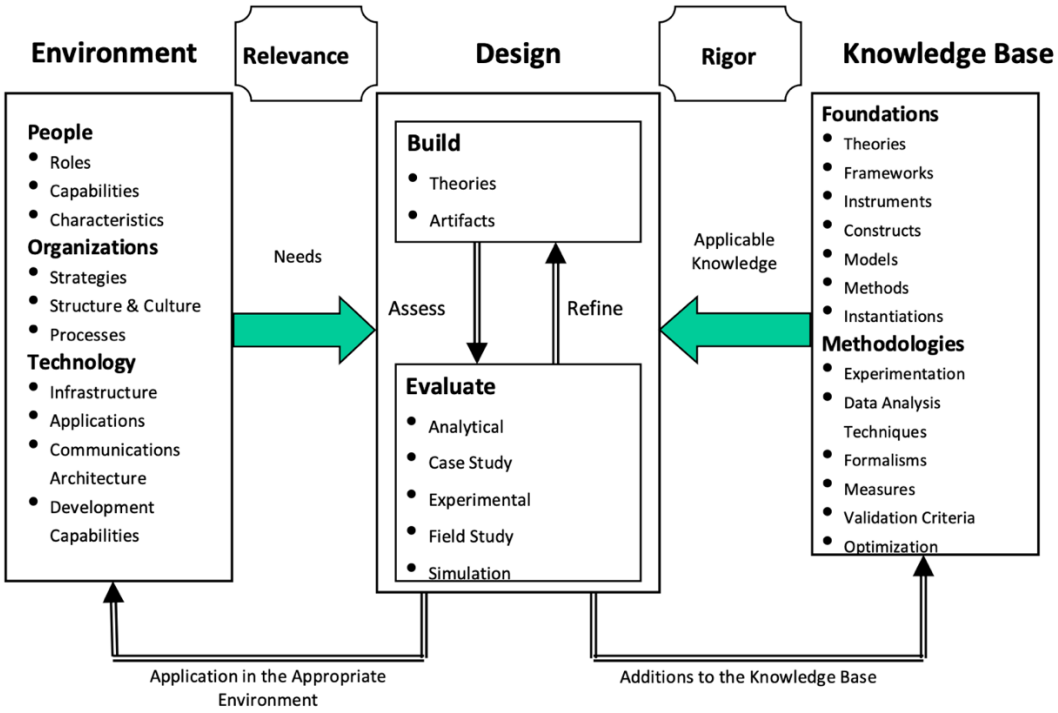


Figure 1: Design Science Research framework by vom Brocke et al. (2020)

The DSR methodology process model as proposed by Pfeffers et al. (2008) provides a commonly used guidance to the procedure in design science research (Figure 2). The model consists of 6 steps: 1) problem identification and motivation, 2) definition of objectives of a solution, 3) design and development of an artifact, 4) demonstration of the use of an artifact, 5) evaluation of the artifact and 6) communication.

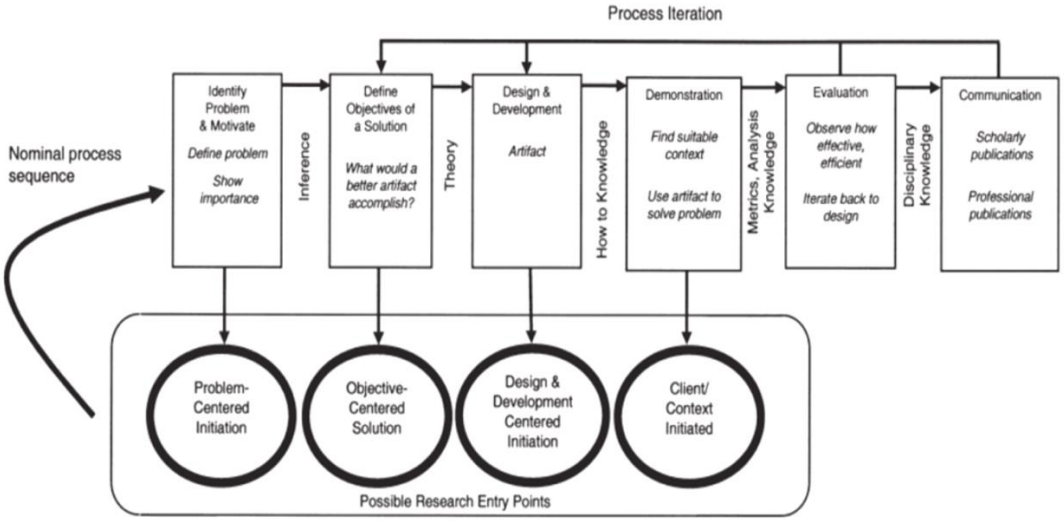


Figure 2: DSR methodology process model by vom Brocke et al. (2020)

The first two papers involve alternative data in tax audits and the new reporting format iXBRL whereas essays 3-6 take a closer look at the use of knowledge graphs and their benefits.

The first essay examines how alternative data can support tax audits. Here, the various sources of alternative data are discussed, as well as various coding examples for their applications.

The second paper deals with the reporting format iXBRL, which is still new in Germany. The quality of the company accounts of SMEs from UK is examined. This is of interest, as iXBRL will also gain importance in Germany due to the introduction of ESEF and the EU taxonomy on sustainability.

The third paper analyzes to what extent the UK iXBRL company accounts can be used with the help of a knowledge graph. The iXBRL company accounts are supplemented by freely accessible alternative data. The creation of an artifact based on a knowledge graph forms the basis for further analyses here. The use of a knowledge graph makes it possible to link the data better, but also to analyze it. This is illustrated in the paper with a couple of examples.

The fourth paper examines the use of a labeled property graph (Neo4j) to support the analysis of connections between boards of directors, supervisory boards, and auditors. Until now, these evaluations have been very difficult and time-consuming. The use of a labeled property graph offers a possibility to simplify the analysis significantly.

The fifth paper highlights the people behind the Wirecard scandal. The focus here is on the people behind the scandal and how it could have come to this. In order to answer this question, the curricula vitae of the members of Wirecard's supervisory board and management board were collected and converted into RDF format. Already during the collection of the data, it was noticeable that there is significantly less information from the persons concerned than from other companies. In order to go into more detail about the differences in the composition, but also the people on the supervisory board and management board of Wirecard, the results are compared with the other DAX30 companies. For this purpose, a knowledge graph based on RDF data is created. This helps with both the analysis and the visualization of the data. This artifact helps analyze and visualize the data and answer the question of what was different at Wirecard's supervisory board and management board compared to the other companies in the DAX30.

The sixth essay complements the prior papers which focus on the benefits of knowledge graphs by making knowledge graphs accessible to a larger group of people. This artifact serves as the basis for the further application and evaluation of an educational game for teaching RDF and SPARQL queries.

2 Motivation and research questions

2.1 Zielführende Betriebsprüfungen durch Nutzung von „Alternative Data“?

The first paper deals with the question of how alternative data sources can be adopted for a target-oriented tax audit. As part of the tax process, the tax authorities must both select companies for audit and identify key tax risk areas within the companies (Haarmann, 1977; Sauer, 1988).

Traditionally, the administration relies on assessment data, comparisons of key figures with other companies and, if necessary, e-balance sheet data (Bittner et al., 2016). The reporting requirements of country-by-country reporting (Lutz & Seebeck, 2020) and the reporting requirements for tax arrangements (EU Mutual Assistance Directive) can also provide clues. In most cases, however, the information provided by the taxpayer is (initially) unaudited data. Its completeness and accuracy of which cannot be granted, especially in the case of "aggressive" taxpayers. Therefore, the supplementary use of other data sources, so-called "alternative data", also appears to make sense (Peemöller & Kregel, 2010).

The availability of alternative data sources on the internet is steadily increasing. Alternative data can be defined here as data that has not been declared by taxpayers during the taxation procedure and that was not already available within the tax administration (Monk et al., 2019).

In the theoretical part, this paper first addresses the question of which approaches to plan tax audits are already common practice both nationally and internationally. Subsequently, various national and international developments are presented. On this basis, the following sections examine how an exploitation of alternative data could alert the tax administration about inconsistencies with tax declarations or sources of risk.

Alternative data has various characteristics that can make it easier or more difficult to utilize. A comprehensive selection of alternative data is identified, analyzed and evaluated in order to show how it could be appropriate in company selection or to set the focus of the tax audit.

Finally, some prototypes show how data queries for selected questions in a tax audit can be implemented. The first step is to clarify the technological basis.

This includes a comparison of different concepts for data storage, linking and retrieval. On this basis, three concrete coding examples are presented.

2.2 Analyzing the quality of iXBRL company accounts in the UK

The implementation of the eXtensible Business Reporting Language (XBRL) has significantly changed the way how companies provide financial information. Now it is possible to evaluate the information of XBRL reports automatically. Gone are the days when annual reports were only available in PDF format and preferably for reading only. iXBRL (Inline XBRL) is the further development of XBRL instance documents. iXBRL documents can be read by machines as well as by humans.

Data is an essential resource for companies. Prior literature shows that the use of XBRL in companies reduces cost (Blankespoor, 2019), increases efficiency (Dhole et al., 2015; Amin et al., 2018; Chen & Zhou, 2019; Du & Wu, 2019), and lowers complexity (Cong et al., 2019; Hoitash & Hoitash, 2018; Li & Nwaeze, 2018). Most of the research is focused on listed companies. However, literature towards the quality of iXBRL company accounts is lacking.

This paper closes the existing gap in literature by investigating the quality of filed iXBRL accounts from small and medium-sized enterprises (SME) in the years 2016–2019 in the UK. In total, I analyze 882.796.471 iXBRL tags from 2.892.841 companies.

The results explain the eight most common errors in iXBRL reports such as a wrong structure of the tags, incorrectly tagged references or errors in the formatting. Overall, the results show, that the quality of the iXBRL accounts is very good and that most iXBRL filings do not include errors. This result is especially interesting. The UK is the only country where the filing of iXBRL accounts of SME companies is mandatory.

In addition, this study focuses on SME companies and proves that the company size does not necessarily determine the quality of the iXBRL filing.

The paper contributes to the existing literature by evaluating the quality of iXBRL reports in more detail. The results are of great relevance with regards to the implementation of the new European Single Electronic Format (ESEF) in the European Union.

2.3 A Knowledge Graph from UK Financial Statements

The third paper discusses how iXBRL company accounts from the UK can be processed to merge with other data. The UK is already working with iXBRL since 2011. In UK's case, not only data from the big, capital market-oriented companies are published, but also from a majority of small and medium sized enterprises (Company Reporting in the UK – an XBRL Success Story, 2015; XBRL UK, 2021). To take further advantage of the data, the aim of this paper is to create a knowledge graph from all the company information available in the iXBRL company accounts. The UK Companies House has a "find and update company information"-register, which provides further information. In addition to that, we will add further information to the knowledge graph. Free linked open data from databases like DBpedia will be added. The knowledge graph allows us to interconnect and verify the information from the different databases and to visualize links within the data (Dadzie & Rowe, 2011).

To implement the paper technically, data from different formats is integrated and converted to the knowledge graph presented in a Resource Description Framework (RDF) format. RDF is a machine-readable format on the semantic web, which helps to integrate multiple sources of data (Ashraf & Hussain, 2012). There are three data formats utilized in this study, which are 1) iXBRL-formatted financial reports from Companies House, 2) other information in Json (JavaScript Object Notation) format from the Companies House e.g., a company profile and an officer detail, and 3) public linked data such as DBpedia.

A RDF mapping language tool and Python script are used to transform the information into the RDF presentation. An ontology is built by a tool called Protégé to provide a semantic relation to the data. Putting this all together, the graph consists of the converted iXBRL reports, the ontology, and the linked data. In the end, SPARQL, a query language for RDF, is used. This allows us not only to evaluate the quality of the knowledge graph, but also to perform analysis.

The paper contributes to the existing literature by using the concept of XBRL and iXBRL and linking it to other data sources. In particular, the feeding of a knowledge graph is new.

With the help of the knowledge graph, data from different sources can not only be linked together, but also analyzed (Shen et al., 2015; Burdick et al., 2011) through the SPARQL query. The queries allow us to perform investigations in a larger data collection that would otherwise not have been possible.

2.4 How to visualize relationships between supervisory board and management board members and auditors using Neo4j

The fourth paper uses Neo4j to analyze relationships between supervisory and management board members of companies listed in the German stock index (DAX30). The topic is of interest for two reasons. First, the special German institutional setting follows a so called "principle of separation". The supervision and the management of corporations should be clearly separated (two-tier system). In the past and present, there are various examples of multiple mandates and personal ties between the members, including changes from the management board to the supervisory board in subsequent years or from auditors to the management (Gröls, 2011; Oehmichen, 2011). Second, those relationships are difficult to identify and to analyze. Frequently used software such as Excel or SQL offer analytical options, but meaningful data visualizations are not easy. Specifically, the connections of members from the management board, the supervisory board and the auditors are hard to visualize with common software. Therefore, we have chosen to use a graph database (Neo4j) to detect, visualize and examine personal ties (Vicknair et al., 2010; Cheng et al., 2019; Gong et al., 2018).

The sample is based on the DAX30 companies in the year 2019. In total, the data includes 480 supervisory board members, 197 management board members and 56 auditors.

This paper shows how Neo4j can be used to visualize personal networks in the German two-tier system. The results reveal that the members of the management and supervisory boards have various personal networks. This might cause problems because the objectivity and independence of supervisory board members and auditors could be reduced, which in turn affects their oversight ability in a negative way. In addition, the possibilities offered by knowledge graphs and Neo4j in particular are demonstrated.

Especially the visualization of the connections offers a big advantage compared to conventional databases.

2.5 What can we learn from Knowledge Graphs? A Wirecard perspective

The fifth paper is dedicated to the Wirecard scandal, which was one of the most shocking economic events in Germany in 2020. The bankruptcy on June 25 of German financial services provider and former DAX30 company Wirecard AG has intensified calls for more comprehensive financial supervision (Krahnen & Langenbucher, 2020; Véron, 2020). Wirecard, which was blamed on accounting irregularities, promoted by a lack of supervision, owes creditors more than €3.5 billion (almost $4 billion). German law requires that members of management boards, supervisory boards and auditors be "independent" of each other. However, this only refers to the status of affairs and does not include long-term relationships (Gröls, 2011; Oehmichen, 2011).

In order to capture these aspects, we created a knowledge graph in our study containing details of each person of interest, such as previous positions and their education (Singhal, 2012). To increase the scope of this information, the graph was enriched with data from external open data sources. This enabled us to gather all relevant information from management and supervisory board members from the 30 largest companies in Germany for the fiscal year 2019, building on previous research in the fourth paper. Relevant information includes place of birth, date of birth, education and work experience. Our knowledge graph contains 745 people, 1.203 companies and organizations, 5.116 roles and 1.128 degrees or educational programs. All this information helps us to understand what was different at Wirecard and what may be the reasons why Wirecard failed.

Using a knowledge graph enables us to automatically detect all kinds of ties between supervisory and management board members, whereas detecting them manually requires more effort and is more error prone. The collection of the data reveals that there was little to no information available for most of the supervisory and management board members of Wirecard, which is very uncommon for a DAX30 firm. For this reason, the question of independence did arise. In addition to that, it is obvious that the management board has far less work experience than managers from comparable companies. When looking at the experience of Wirecard’s supervisory board members, it becomes clear they had little or no familiarity with board activities.

Furthermore, our graph shows that supervisory board members from Wirecard are less connected to other members of supervisory boards compared to other DAX30 companies.

The paper contributes to the existing literature by using a new technology, knowledge graphs, to gain deeper insights about an existing problem. Knowledge graphs can be used not only to uncover connections between persons of interest and thus clarify issues of independence, but also to better explore fundamental expertise and experience of members of management and supervisory boards.

2.6 WireGraph - Entwicklung eines Lernspiels mit Prosumentenumgebung zur Förderung der fachlichen und digitalen Kompetenz von Wirtschaftsstudierenden

The last paper deals with the teaching of digital competencies. The European Commission warns that insufficient attention is being paid to the development of digitalization in university teaching. Many graduates lack digital competence. On the other hand, the digital requirements on the labor market are constantly growing. This also applies to professions in the financial sector (Henselmann & Scherr, 2017). The reference model of digital competence proposed by the European Commission is suitable for incorporating digital competence into teaching (Europäische Kommission, 2017). However, there is a great need for research on how the reference model can be explicitly transferred into teaching.

In this paper, we present the concept of the WireGraph learning game, which students of business studies use to learn how to work with knowledge graphs in order to work through the Wirecard scandal. The learning game teaches various competencies. On the one hand, business students are encouraged to deal with knowledge graphs, RDF and SPARQL queries. Only when students have familiarized themselves with these concepts will they be able to solve the game. On the other hand, the Wirecard case will also be discussed in more detail. The students will get to know the history of the scandal but also the people behind the scandal. Thus, even students who already have advanced IT knowledge can learn something about corporate governance in the game.

However, students should also be encouraged to contribute their own ideas (Sicilia et al., 2018). For this purpose, a prosumer environment is provided, with which the competence area of digital content creation of the reference model is taught completely and

across all performance levels. Students not only learn the basics of RDF and SPARQL but can also use them specifically for their ideas and further expertise.

The conception of the WireGraph learning game is the first stage of a multi-stage project. The following steps are the elaboration of the learning game and subsequently the application and the evaluation of the didactic success. This paper with its presentation of the initial situation, the goals and the chosen approach forms the basis for the following elaboration of the learning game. The described research questions are essential for the later evaluation of the didactic success.

3 Structure of the thesis

This doctoral thesis is structured as follows. Chapter A (Introduction) provides the background and the motivation and research questions of the six papers. Chapter B (Main part) contains the six paper and seeks to answer the research questions:

- Section B.1 provides the first research paper “Zielführende Betriebsprüfungen durch Nutzung von „Alternative Data“?“.
- Section B.2 contains the second research paper titled “Analyzing the quality of iXBRL company accounts in the UK”.
- Section B.3 includes the third paper “A Knowledge Graph from UK Financial Statements”.
- Section B.4 provides the fourth paper “How to visualize relationships between supervisory board and management board members and auditors using Neo4j”.
- Section B.5 contains the fifth paper “What can we learn from Knowledge Graphs? A Wirecard perspective”.
- Section B.6 inludes the sixth paper “WireGraph - Entwicklung eines Lernspiels mit Prosumentenumgebung zur Förderung der fachlichen und digitalen Kompetenz von Wirtschaftsstudierenden“.

Finally, chapter C (Conclusions) concludes the doctoral thesis with a brief summary of the main findings, the limitations and future research avenues.

References

Ashraf, J., & Hussain, O. K. (2012). Integrating Financial Data Using Semantic Web for Improved Visibility. In *Proceedings of 8th International Conference on Semantics, Knowledge and Grids (SKG)* (pp. 265-268). Beijing, China: IEEE.

Barrasa, Jesús. (2017, August 18). RDF Triple Stores vs. Labeled Property Graphs: What's the Difference?. Neo4j. https://neo4j.com/blog/rdf-triple-store-vs-labeled-property-graph-difference/.

Becker, J., Niehaves, B., Obrich, S. & Pfeiffer, D. (2009). Forschungsmethodik einer Integrationsdisziplin – Eine Fortführung und Ergänzung zu Lutz Heinrichs „Beitrag zur Geschichte der Wirtschaftsinformatik" aus gestaltungsorientierter Perspektive. In J. Becker, H. Krcmar & B. Niehaves (Eds.) *Wissenschaftstheorie und gestaltungsorientierte Wirtschaftsinformatik* (pp. 1-22). Physica. https://rd.springer.com/chapter/10.1007/978-3-7908-2336-3_1.

Bittner, T., Dawid, R., & Metzner, S. (2016): Typische Problemfelder in Betriebsprüfungen. In D. Roman (Ed.) *Verrechnungspreise* (pp. 243-276). Springer Gabler. https://doi.org/10.1007/978-3-658-09377-8_6.

Burdick, D., Hernandez, M., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L., Stanoi, I., Vaithyanathan, S., & Das, S. (2015). Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study. *SSRN Electronic Journal*. https://dx.doi.org/10.2139/ssrn.2666384.

Cheng, Y., Ding, P., Wang, T., Lu, W., & Du, X. (2019). Which Category Is Better: Benchmarking Relational and Graph Database Management Systems. *Data Science and Engineering*, 4(4), 309-322.

Dadzie, A. S., and Rowe, M. (2011). Approaches to visualising Linked Data: A survey, *SemanticWeb*, 2(2), 89–124. 10.3233/SW-2011-0037.

Europäische Kommission. (2017). Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen- über eine europäische Erneuerungsagenda für die Hochschulbildung. COM, 247.

Fensel, D., Simsek. U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I, Umbrich, J. & Wahler, A. (2020). Introduction: What Is a Knowledge Graph? In *Knowledge Graphs*. Springer. https://doi.org/10.1007/978-3-030-37439-6_1.

Gartner. (2020, July). Hype Cycle for Artificial Intelligence, 2020. https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020.

Gong, F., Ma, Y., Gong, W., Li, X., Li, C., & Yuan, X. (2018). Neo4j graph database realizes efficient storage performance of oilfield ontology. PLOS ONE. 13(11). Gröls, M. 2011. Die letzten Herren der "Deutschland AG"? *Der Aufsichtsrat*: (07-08): 106-107.

Gröls, M. (2011). Die letzten Herren der "Deutschland AG"? *Der Aufsichtsrat*: (07-08): 106-107.

Haarmann, W. (1977): Abgabenordnung - Einführung und Begriffsbestimmungen, Inf. 1977.

Henselmann, K., & Scherr, E. (2017). Auswirkungen der Digitalisierung auf den Berufsnachwuchs in der Wirtschaftsprüfung. In J. Baldauf and Graschitz, S. Eds.) *Theorie und Praxis aus Rechnungswesen und Wirtschaftsprüfung* (pp. 231–24). LexisNexis Verlag.

Herren Lee, A. (2020, November 17). The Promise of Structured Data: True Modernization of Disclosure Effectiveness. U.S. Securities and Exchange Commission. https://www.sec.gov/news/speech/lee-structured-data-2020-11-17.

Hevner, A., March, S., Park, J. & Ram, S. U (2004). Design Science in Information Systems Re-search. *MIS Quarterly*.

Krahnen, J. P., & K. Langenbucher, K. (2020). The Wirecard lessons: A reform proposal for the supervision of securities markets in Europe. SAFE Policy Letter, Research Report. (88). https://www.econstor.eu/handle/10419/222230.

Lutz, F., & Seebeck, A. (2020). OECD veröffentlicht aktualisierte Guidance zum Country-by-Country Reporting, *Internationales Steuerrecht (IStR) 29/2020*, 55-59.

Metinko, C. (2021, June 17). Neo4j Hits $2B-Plus Valuation After $325M Raise. Crunchbase News. https://news.crunchbase.com/news/neo4j-hits-2b-plus-valuation-after-325m-raise/#:~:text=San%20Mateo%2C%20California%2Dbased%20Neo4j,for%20the%20graph%20database%20company.

Monk, A., Prins, M., & Rook, D. (2019). Rethinking Alternative Date in Institutional Investmet. *The Journal of Financial Data Science*. 14-31. https://doi.org/10.2139/ssrn.3193805.

Oehmichen, J. (2011). Mehrfachmandate von Aufsichtsratsmitgliedern: Eine Panel-Analyse ihrer Wirkung in deutschen Unternehmen. In Lindstädt (Ed.), *Schriften zu Management, Organisation und Information*, Augsburg, Germany: Rainer Hampp Verlag.

Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., Krcmar, H., Loos, P., Mertens, P., Oberweis, A. & Sinz, E.J. (2011). Memorandum on design-oriented information systems research. *European Journal of Information Systems*. 20. 7-10. https://link.springer.com/article/10.1057/ejis.2010.55.

Peemöller, V. H., & Kregel, J. (2010). *Grundlagen der internen Revision*. Erich Schmidt Verlag.

Peffers, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*. 24. 45-77. https://doi.org/10.2753/MIS0742-1222240302.

RDF. (2014, February 25). RDF Working Group. https://www.w3.org/RDF/.

Robinson, I., Webber, J., & Eifrem, E. (2013). *Graph databases*. O'Reilly Media, Inc.

Sauer, O. (1988). Steuerliche Außenprüfung. Vahlen Verlag.

Schreiner, M. W. 1., Hess, T. 1., & Benlian, A. 1. (2015). *Gestaltungsorientierter Kern oder Tendenz zur Empirie?* Ludwig-Maximilians-Univ., Inst. f. Wirtschaftsinformatik u. Neue Medien. https://katalog.ub.tu-braunschweig.de/vufind/Search2Record/821607464.

Shen, W., Wang, J., & Han J. (2015). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2), 443–460. 10.1109/TKDE.2014.2327028.

Sicilia, M., Barriocanal, E.G., Sánchez-Alonso, S., Rózewski, P., Kieruzel, M., Lipczynski, T., Royo, C., Uras, F., & Hamill, C. (2018). Digital skills training in Higher Education: insights about the perceptions of different stakeholders. *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*. 781-787. https://doi.org/10.1145/3284179.3284312.

Singhal, A. (2012). Introducing the Knowledge Graph: things, not strings. Google. https://blog.google/products/search/introducing-knowledge-graph-things-not/.

Véron, N. (2020). The Wirecard debacle calls for a rethink of EU, not just German, financial reporting supervision. Bruegel. https://www.bruegel.org/2020/06/the-wirecard-debacle-calls-for-a-rethink-of-eu-not-just-german-financial-reporting-supervision/.

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., & Wilkins, D. (2010). A comparison of a graph database and a relational database: a data provenance perspective. *Proceedings of the 48th annual southeast regional conference*. 1-6.

Vom Brocke J., Hevner A. & Maedche A. (2020). Introduction to Design Science Research. In: vom Brocke J., Hevner A., Maedche A. (Eds.) *Design Science Research. Cases*. Cham. https://doi.org/10.1007/978-3-030-46781-4_1.

Wilde, T. & Hess, T., (2007). Forschungsmethoden der Wirtschaftsinformatik. *Wirtschaftsinformatik*. 49(4). Springer. (pp. 280-287). 10.1007/s11576-007-0064-z.

XBRL UK. (2021). XBLR in the UK. https://www.xbrl.org.uk/projects/.

XBRL UK. (2015). Company Reporting in the UK – an XBRL Success Story. https://www.xbrl.org.uk/resources/whitepapers/UKcompanyReporting-XBRL-v1.pdf.

Chapter B

Main Part

Contents – Chapter B

Section B.1

Zielführende Betriebsprüfungen durch Nutzung von „Alternative Data“?

(with Klaus Henselmann & Andreas Seebeck)

Abschlussbericht zum Forschungsprojekt der Prof. Dr. oec. Westerfelhaus Stiftung

Veröffentlicht unter:

https://www.econstor.eu/handle/10419/248286/

(die Zitierweise mit Fußnoten wurde beibehalten)

Contents – Section B.1

Zielführende Betriebsprüfungen durch Nutzung von „Alternative Data“?

Abstract

Im Rahmen des Steuerverfahrens muss die Finanzverwaltung sowohl Unternehmen für die Betriebsprüfung auswählen als auch innerhalb der Unternehmen steuerliche Risikoschwerpunktbereiche identifizieren. Traditionell stützt die Verwaltung sich dabei auf Veranlagungsdaten, Kennzahlenvergleiche mit anderen Unternehmen sowie ggf. auf E-Bilanz-Daten. Anhaltspunkte können auch die Meldepflichten des sog. Country-by-Country-Reporting und die Meldepflichten für Steuergestaltungen (EU-Amtshilferichtlinie) bieten. Meist handelt es sich bei Angaben des Steuerpflichtigen jedoch um (zunächst) ungeprüfte Daten. Sinnvoll erscheint somit auch die ergänzende Verwendung anderer Datenquellen, sog. "Alternative Data".

Auf dieser Basis wird in den folgenden Abschnitten abgeleitet, wie eine Erschließung alternativer Daten die Finanzverwaltung auf Unstimmigkeiten zur Steuerdeklaration oder besondere Risikoquellen hinweisen könnte.

Keywords

Besteuerungsverfahren; Außenprüfung; Risikoanalyse; Open Data; Linked Open Data; RDF; SPARQL; Graph Data Base

1 Einführung

Im Rahmen des Steuerverfahrens muss die Finanzverwaltung sowohl Unternehmen für die Betriebsprüfung auswählen als auch innerhalb der Unternehmen steuerliche Risikoschwerpunktbereiche identifizieren.

Traditionell stützt die Verwaltung sich dabei auf Veranlagungsdaten, Kennzahlenvergleiche mit anderen Unternehmen sowie ggf. auf E-Bilanz-Daten. Anhaltspunkte können auch die Meldepflichten des sog. Country-by-Country-Reporting (§§ 138a ff. AO) und die Meldepflichten für Steuergestaltungen (EU-Amtshilferichtlinie) bieten.

Meist handelt es sich bei Angaben des Steuerpflichtigen jedoch um (zunächst) ungeprüfte Daten, deren Vollständigkeit und Richtigkeit gerade bei „aggressiven" Steuerpflichtigen nicht pauschal unterstellt werden kann.

Sinnvoll erscheint somit auch die ergänzende Verwendung anderer Datenquellen, sog. „Alternative Data". Die Verfügbarkeit alternativer Datenquellen im Internet steigt stetig an. Der Begriff „Alternative Data" stammt eigentlich aus der Finanzanalyse.[1] Er umfasst dort alle nicht-konventionellen Datenquellen, insbesondere solche, die über Börsenhandelsdaten und Daten aufgrund der Finanzberichterstattung durch das Unternehmens selbst hinausgehen. Alternative Daten werden bei der Finanzanalyse verwendet, um die vom Unternehmen gemachten Aussagen zu ergänzen, zu bestätigen oder eben zu widerlegen. Diesen Grundgedanken kann man auch auf die Besteuerung übertragen. Als Alternative Daten können hier solche Daten bezeichnet werden, die nicht im Zuge des Besteuerungsverfahrens von Steuerpflichtigen erklärt oder durch deren Verarbeitung innerhalb der Finanzverwaltung geschaffen wurden.

Der vorliegende Forschungsbericht geht im Grundlagenteil zunächst der Frage nach, welche Ansätze zur Planung der Betriebsprüfung national und international bereits üblich sind. Ausgehend von Aufgaben, Zielen und Anlässe der klassischen Betriebsprüfung wird hier auf die anlassabhängige und anlassunabhängige Betriebsprüfung und die traditionellen Methoden der Prüfungsplanung eingegangen.

Anschließend werden verschiedene nationale und internationale Entwicklungen vorgestellt, die unterstützend bei der Betriebsprüfungsplanung und -durchführung eingesetzt werden können.

[1] Vgl. Monk et al. (2019).

Die verschiedenen Projekte befassen sich u.a. mit der Sammlung und Auswertung von Daten zur Aufdeckung von Betrug, Geldwäsche o.ä. und weisen vielfach eine grundsätzliche Eignung zu Betriebsprüfungszwecken auf.

Auf dieser Basis wird in den folgenden Abschnitten abgeleitet, wie eine Erschließung alternativer Daten die Finanzverwaltung auf Unstimmigkeiten zur Steuerdeklaration oder besondere Risikoquellen hinweisen könnte. Alternative Data kann somit nicht nur die risikoorientierte Unternehmensauswahl verbessern, sondern auch die Betriebsprüfungen zielgerichteter und damit möglicherweise schneller und zugleich intensiver machen.

Im Hauptteil geht es ferner um eine Bestandsaufnahme Alternativer Daten. Alternative Daten haben verschiedene Eigenschaften, die ihre Nutzung erleichtern oder erschweren können. Es wird eine umfassende Auswahl von Alternative Data identifiziert, analysiert und beurteilt, um aufzuzeigen, welche dieser Daten bei der Unternehmensauswahl bzw. Schwerpunktsetzung im Zuge der Betriebsprüfung dienlich sein können.

Abschließend wird im Hauptteil prototypisch gezeigt, wie konkrete Datenabfragen für ausgewählte Fragestellungen im Kontext der Betriebsprüfung aussehen könnten. Dazu erfolgt zunächst ein Vergleich von unterschiedlichen Konzepten der Datenspeicherung, -verknüpfung und -abfrage. Auf dieser Grundlage wird ein Umsetzungsvorschlag unterbreitet.

Eine Zusammenfassung wichtiger Erkenntnisse und ein Ausblick schließen die Arbeit ab.

Einige Probleme sollen oder können nicht Gegenstand der vorliegenden Untersuchung sein. Hierunter fällt erstens die juristische Frage, inwieweit die vorgestellten Konzepte für Datenabfragen und/oder Datensammlungen nach heutigem Recht bereits zulässig wären. Ob und gegebenenfalls, wie hierfür der rechtliche Rahmen geändert werden müsste, soll an dieser Stelle offenbleiben. Die Erfahrung zeigt jedoch, dass das Recht nicht unveränderlich ist, sondern Maßnahmen gegen „Base Erosion and Profit Shifting“ (BEPS)[2] zu entsprechenden Anpassungen führen können.[3]

[2] Vgl. OECD (2021).

[3] Beispielsweise wird das von der Europäischen Kommission vorgeschlagene „*öffentliche* Country-by-Country-Reporting“ zumindest über bestimmte Länder (Staaten der EU sowie Steueroasen gemäß EU-Liste) künftig Pflicht für große internationale Konzerne. So die Entscheidung der EU-

Daneben werden auch praktische Fragen der Implementierung ausgeklammert. Hierunter fallen etwa die organisatorische Verankerung innerhalb der Finanzverwaltung, Anforderungen an die Technik und die Qualifikation der Mitarbeiter, das erforderliche Budget sowie dessen Finanzierung.

In diesem Zusammenhang soll auch nicht hinterfragt werden, wie die innerhalb der Finanzverwaltung bereits vorhandenen Daten durch bessere Auswertungen, umfangreichere Verknüpfungen und vermehrte Abfragen bereits Hinweise für zielgerechte Betriebsprüfungen liefern könnten.[4]

Last but not least verändern sich das Angebot an Alternativen Daten sowie die Bezugsmöglichkeiten ständig. Der im nachfolgenden Text dargestellte Stand bezieht sich grundsätzlich auf die Situation im September 2021.

2 Grundlagen

2.1 Klassische Betriebsprüfung

2.1.1 Aufgabe

Der Gesetzgeber der AO hat an Stelle des früheren Begriffs „Betriebsprüfung" bereits im Jahr 1977 den Begriff „Außenprüfung" gesetzt. Dadurch kommt klar zum Ausdruck, dass die Betriebsprüfung weit auszulegen ist und nicht nur steuerliche Verhältnisse von Betrieben, sondern mitunter auch von Privatpersonen erfassen soll.[5] Dies erleichtert der Finanzverwaltung den Zugang zu umfassenden Daten, die mitunter außerhalb der betrieblichen Sphäre der Unternehmen liegen. Die Finanzverwaltung selbst behält den Begriff Betriebsprüfung bei, siehe etwa die „Allgemeine Verwaltungsvorschrift für die Betriebsprüfung – Betriebsprüfungsordnung (BpO)", weshalb auch im Rahmen dieses Forschungsberichts die Begriffe Betriebsprüfung und Außenprüfung synonym verwendet werden.

Eine Betriebsprüfung kann der umfassenden Aufklärung besteuerungsrelevanter Sachverhalte dienen, die im Veranlagungsverfahren von Amts wegen nicht oder nicht

Wirtschaftsminister am 25.2.2021. Das ist bemerkenswert, denn es handelt sich um Daten, die bisher dem Steuergeheimnis unterlagen. Vgl. Kafsak / Schäfers (2021).

[4] Hinweise auf das große Potenzial gibt ein Bericht des Bundesrechnungshofs zur Verbesserung der Umsatzsteuerbetrugsbekämpfung. Bundesrechnungshof (2020).

[5] Vgl. Haarmann (1977), S. 135; sowie Sauer (1988), S. 21f.

zweckmäßig überprüft werden können. Sie ist ein Instrument der umfassenden finanzbehördlichen Sachaufklärung, welches letztlich die Steuergerechtigkeit fördert.[6]

In Abhängigkeit von der Größe des Unternehmens beträgt der **Prüfungsturnus** alle drei bis 20 Jahre. Während Großunternehmen im Bundesdurchschnitt alle drei bis vier Jahre einer Betriebsprüfung unterzogen werden, kommen Kleinbetriebe nur etwa alle 15 bis 20 Jahre in den Genuss einer Prüfung. Zu beachten ist, dass es erhebliche regionale Unterschiede gibt, die nicht zuletzt durch die Personalsituation in den jeweiligen Verwaltungseinheiten bedingt sind.

Der **Prüfungszeitraum** umfasst i.d.R. drei Jahre (§ 4 BpO), kann sich aber in begründeten Fällen, in denen mit nicht unerheblichen Änderungen der Besteuerungsgrundlage zu rechnen ist oder bei Verdacht einer Steuerstraftat oder -ordnungswidrigkeit, ausgeweitet werden. Eine Ausdehnung des Prüfungszeitraums ist dabei nicht bereits darin zu sehen, dass sich ein Prüfer Urkunden außerhalb des Prüfungszeitraums vorlegen lässt, soweit er diese für eine Schätzung für die geprüften Jahre benötigt und dies entsprechend begründen kann.[7]

Prüfungsgegenstand kann der gesamte für die Entstehung und Ausgestaltung eines Steueranspruchs erhebliche Sachverhalt sein – sowohl der Höhe als auch dem Grunde nach. Es werden die erklärten Einkünfte des Unternehmers vollumfänglich überprüft. Zu diesem Zwecke werden alle steuerlich relevanten Sachverhalte gewürdigt. Hierunter fallen alle Einkunftsarten und Besteuerungsmerkmale, selbst wenn sie mit den betrieblichen Verhältnissen in keinem direkten Zusammenhang stehen. [8]

2.1.2 Ziele und Anlässe

Die Betriebsprüfung ist ein spezielles Verwaltungsverfahren, mit dem die Finanzbehörde das ihr in § 85 AO auferlegte **Ziel** durchsetzen kann: Die Steuer nach gesetzlichen Maßgaben gleichmäßig festzusetzen und zu erheben. Übereinstimmend mit dem Gesetzeswortlaut der AO charakterisiert der BFH die Betriebsprüfung als ein Instrument zur Sicherstellung der Gleichmäßigkeit der Besteuerung.[9]

---

[6] Vgl. BFH v. 9.8.1991, BStBl 1992 II S. 220; Tipke/Lang § 22 Rz. 225.

[7] Vgl. BFH v. 4.2.1988 – V R 57/83, BStBl 1988 II S. 413.

[8] Vgl. Harle / Nüdling / Olles (2020), S. 9f.

[9] Vgl. BFH v. 28.9.2011 – VIII R 8/09, BStBl 2012 II S. 395 = BB 2012 S. 3 737.

Die Betriebsprüfung hat sich auf die tatsächlichen und rechtlichen Verhältnisse, die für die Steuerpflicht und die Steuerbemessung maßgebend sind – zugunsten wie zuungunsten des Steuerpflichtigen – zu erstrecken (§ 199 Abs. 1 AO).

Die Betriebsprüfung dient letztlich dazu, die Wettbewerbsneutralität der Besteuerung sicherzustellen. Sie ist darauf ausgerichtet die steuerlichen Verhältnisse des Steuerpflichtigen zu überprüfen und umfasst eine oder mehrere Steuerarten, wie Einkommen- und Körperschaftsteuer, Umsatzsteuer und ggf. Gewerbesteuer.

Es kann unterschieden werden zwischen **anlassunabhängigen** und **anlassabhängigen** Außenprüfungen. Die Entscheidung, ob und wenn ja, wann bei einem Betrieb eine Betriebsprüfung durchgeführt wird, liegt gemäß § 5 AO im pflichtgemäßen Ermessen der Finanzbehörde. Bei der Entscheidung ist der verfassungsrechtliche Verhältnismäßigkeitsgrundsatz zu beachten.

Eine besondere Begründung der Prüfungsanordnung oder ein konkreter Anlass als Begründung für die Prüfung ist aufgrund der Rechtsgrundlage des § 193 AO nicht erforderlich. So ist eine weitgehende Begründung gemäß BFH-Urteil vom 12.8.2002 nur dann erforderlich, wenn dies zum Verständnis der Prüfungsanordnung erforderlich ist, wie z.B. bei der Erweiterung des Prüfungszeitraums nach § 4 Abs. 3 BpO.

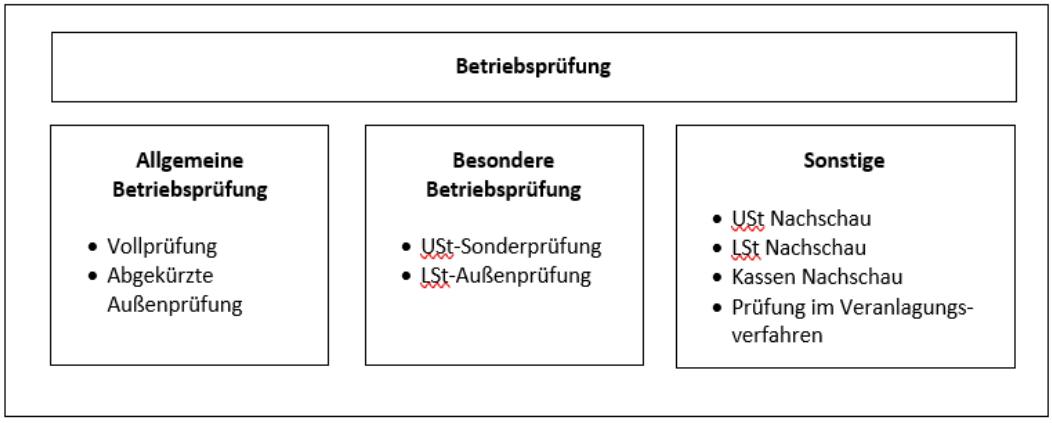


Abb. 1: Durchführungsmöglichkeiten der Außenprüfung
Quelle: Eigene Darstellung in Anlehnung an Harle / Nüdling / Olles (2020), S. 6.

Die **normale Außenprüfung (Vollprüfung)** ist in § 193 ff. AO geregelt. Es gelten die oben dargestellten Rahmenbedingungen. Eine Schwerpunktbildung nach § 7 BpO ist überdies möglich. Daneben besteht die Möglichkeit der zeitnahen Betriebsprüfung nach § 4a BpO.

Sie verfolgt das Ziel möglichst gegenwartsnahe Prüfungszeiträume zu untersuchen und ist häufig durch die Absicht der Verkürzung der Prüfungsdauer bei Großbetrieben oder der Angleichung von Prüfungszeiträumen durch die Finanzämter motiviert. Die wirtschaftlichen Vorteile einer zeitnahen Betriebsprüfung liegen auf der Hand. Neben erhöhter Rechts- und Planungssicherheit für die Unternehmen werden insbesondere die Risiken antizyklischer Steuernachforderungen (in ertragsschwachen Jahren) sowie hohe Nachzahlungszinsen vermieden.

Die **abgekürzte Außenprüfung** (§ 203 AO) beschränkt sich auf die wesentliche Besteuerungsgrundlage. Eine rasche Prüfungsdurchführung wird durch die Beschränkung auf bestimmte Sachverhalte, Zeiträume und Steuerarten ermöglicht. In der Regel bezieht sich die abgekürzte Außenprüfung auf einzelne steuerliche Sachverhalte, die sich auf klar begrenzbare Prüffelder eingrenzen lassen. Die wesentlichen Vorschriften der §§ 193 ff. AO finden Anwendung, mit der Ausnahme, dass kein Anspruch auf Durchführung einer Schlussbesprechung nach § 201 Abs. 1 und § 202 Abs. 2 AO besteht.[10]

Der Zweck von **Sonderprüfungen/Nachschauen** besteht in der Prüfung einzelner Steuerarten (z.B. Umsatzsteuersonderprüfung, Lohnsteuer-Außenprüfung). Es gelten grundsätzlich die allgemeinen Grundsätze für die Betriebsprüfung nach § 193 AO.

Die **Umsatzsteuersonderprüfung** ist eine besondere Form der Betriebsprüfung. Sie wird aufgrund besonderer Umstände des Einzelfalls angeordnet und erstreckt sich ausschließlich auf umsatzsteuerrelevante Sachverhalte. Eine Besonderheit ist, dass der Prüfungszeitraum nicht nach § 4 Abs. 3 BpO auf drei Jahre beschränkt ist.[11]

Die **Lohnsteueraußenprüfung** ist in § 42 ff. EStG geregelt. Prüfungsgegenstand ist die Einbehaltung und Abführung der Lohnsteuer durch den Arbeitgeber. Die Lohnsteuer-Außenprüfung wird durch Verwaltungsakt angeordnet. Für die Mitwirkungspflicht des Arbeitgebers bei der Außenprüfung gilt § 200 AO.

Die **Umsatzsteuernachschau** nach § 27b UStG ist keine Betriebsprüfung im Sinne des § 193 AO. Es handelt sich vielmehr um ein besonderes Verfahren zur zeitnahen Aufklärung möglicher steuererheblicher Sachverhalte.

[10] Vgl. BFH v. 25.1.1989 – X R 158/87, BStBl 1989 II S. 483.

[11] Vgl. FG Rheinland-Pfalz v. 25.5.2007 – 6 K 1325/06, n.v.

Sie findet unangekündigt statt und muss nicht zwingend vom Betriebsprüfer selbst durchgeführt werden. Von der Umsatzsteuernachschau betroffene Unternehmen und Personen haben den Prüfern relevante Unterlagen vorzulegen.

Die ebenfalls unangekündigte **Lohnsteuernachschau** nach § 42g EStG dient der zeitnahen Aufklärung von lohnsteuerrelevanten Sachverhalten. Im Falle von erheblichen Feststellungen kann ohne vorherige Prüfungsanordnung in eine Lohnsteuer-Außenprüfung übergegangen werden.

Die **Kassen-Nachschau** nach § 146b AO wurde mit dem 1.1.2018 eingeführt und ist eine weitere Form der unangekündigten zeitnahen Prüfung zur Sicherung des Steueraufkommens. Dabei werden die Ordnungsmäßigkeit der Kassenaufzeichnungen und die korrekte Übernahme der Kassenaufzeichnungen in die Buchführung überprüft. Auch hier kann im Falle von erheblichen Feststellungen ohne vorherige Prüfungsanordnung zur Außenprüfung übergegangen werden.

Die **Prüfung im Veranlagungsverfahren** findet entweder in Form von Inaugenscheinnahmen gem. § 99 AO oder durch Einzelermittlungen nach §§ 85, 88 und 90 AO statt.

2.1.3 Traditionelle Methoden zur Planung der Betriebsprüfung

Traditionell erfolgt die Prüfungsauswahl bei Steuerpflichtigen, die von § 193 Abs. 1 AO erfasst sind, u.a. nach den folgenden Kriterien:

- Größenklasse
- Zeitlicher Turnus (Routine)
- Auswertung von Betriebsdaten i.e.S.
- Branchenzugehörigkeit
- Allgemeines Risikoprofil
- Zufallskriterien (Generalprävention und Überraschungsmoment).[12]

Das allgemeine Risikoprofil ergibt sich aus verschiedenen Risikoindikatoren. Diese beinhalten beispielsweise hohe Umsatz- oder Betriebsausgabenschwankungen in den

[12] Vgl. Harle / Nüdling / Olles (2020), S. 41.

letzten drei Jahren oder außergewöhnlich hohe Betriebsausgaben im Branchenvergleich.[13] Daneben berücksichtigt die Finanzverwaltung in ihrer Prüfungsplanung aus steuerlicher Sicht typischerweise relevante Sachverhalte wie Betriebsaufspaltungen und -verpachtungen, Abgänge von Wirtschaftsgütern, die regelmäßig stille Reserven beinhalten, und Verträge mit Angehörigen sowie Privatentnahmen oberhalb des Gewinns. Hohe Forderungsverluste, Rückstellungen sowie besonders niedrige oder hohe Rechnungsabgrenzungsposten finden ebenfalls Berücksichtigung bei der Prüfungsauswahl.

Zu einer guten Prüfungsplanung gehört neben dem Studium der Steuerakten, die den Prüfern bereits im Vorfeld über die Unternehmen vorliegen, dass sich die Prüfer umfassend über das zu prüfende Unternehmen informieren. Dazu werden u.a. Unternehmenswebsites, Branchenreports und lokale Medien studiert. Daneben stehen den Finanzbehörden aus verschiedenen Projekten, die im Folgenden kurz vorgestellt werden, potenziell Daten zur Verfügung, die zu Prüfungszwecken genutzt werden können.

2.2 Aktuelle Entwicklungen

2.2.1 Grundsätzliches

Begrenzte Personalkapazitäten der Finanzverwaltung und Effizienzaspekte gehen einher mit dem Erfordernis Betriebsprüfungen zielorientiert, d.h. risikoorientiert, durchzuführen. Dies betrifft die risikoorientierte Auswahl der zu prüfenden Unternehmen sowie die risikoorientierte Prüfungsdurchführung.

Von der Finanzverwaltung wird (halb-)jährlich ein nicht öffentlicher Prüfungsgeschäftsplan nach § 34 Abs. 1 BpO erstellt, der die zu prüfenden Betriebe enthält. Die Qualität der darin festgelegten Prüfungen hängt von verschiedenen Faktoren wie einer gezielten Fallauswahl, einer zutreffenden Schwerpunktbildung, einer rationellen Prüfungsplanung und -organisation und einer zielführenden Prüfungsmethodik (insb. Stichprobenauswahl) sowie dem sinnvollen Einsatz von IT ab. Eine hohe Prüfungsintensität (Quantität) ist dabei nicht automatisch gleichzusetzen mit einer hohen Qualität. Entscheidend für die Qualität der Prüfung ist eine geeignete Fallauswahl und Schwerpunktbildung.

[13] Vgl. Bittner / Dawid / Metzner (2016), S. 243.

Verschiedene nationale und internationale Projekte, die ursprünglich zur Aufdeckung von Betrug, Geldwäsche o.ä. gedacht waren, wurden in den vergangenen Jahren initiiert und umgesetzt. Diese können den Finanzbehörden bei der Fallauswahl, Schwerpunktbildung und rationellen Prüfungsplanung und -organisation behilflich sein. Eine Fülle an Daten wird im Rahmen der Projekte gesammelt und ausgewertet. Einige von ihnen weisen eine grundsätzliche Eignung zu Prüfungsplanungszwecken und zur Durchführung von Außenprüfungen auf.

Im Folgenden werden einige ausgewählte Projekte vorgestellt. Dabei wird neben dem regulatorischen Hintergrund und den zugrundeliegenden Ideen auch auf die jeweils erzeugten Datenbestände eingegangen.

2.2.2 Country-by-Country Reporting

2.2.2.1 Regulatorischer Hintergrund und Idee

Das Country-by-Country Reporting wurde im Jahr 2016 eingeführt, um den Finanzbehörden eine erste Einschätzung von BEPS- und Verrechnungspreisrisiken zu ermöglichen.[14] Die Analyse der CbCR-Daten durch die Finanzbehörden dient nicht dazu Betriebsprüfungen zu ersetzen und kann auch nicht als alleinige Grundlage für Verrechnungspreisanpassungen herangezogen werden. Aber die CbC-Reports erlauben eine erste überschlägige Risikoabschätzung anhand der Risikoindikatoren und unterstützen damit den zielgerichteten Einsatz der begrenzten Ressourcen der Finanzbehörden.[15]

So zielt eine erfolgreiche erste Risikoeinschätzung seitens der Finanzbehörden auf die Aussteuerung tatsächlich risikoreicher Unternehmen zur weiteren Nachforschung und auf eine korrekte Einstufung risikoarmer Unternehmen.[16]

Das Ziel des CbCR besteht darin, ggf. in Verbindung mit weiteren Informationsquellen Verrechnungspreis- und BEPS-Risiken überschlägig einzuschätzen.[17] Die Interpretation der im Rahmen des CbCR berichteten Kennzahlen bzw. Risikoindikatoren erfordert regelmäßig die Gegenüberstellung mit Referenzwerten, wie Marktdaten, festen

[14] Vgl. Lutz / Seebeck (2020), S. 55-59.

[15] Vgl. Lutz / Seebeck (2019b), S. 438-449.

[16] Vgl. Art. 16 Abs. 6 S. 2, 3 der Richtlinie 2011/16/EU (EU-Amtshilferichtlinie); BT-Drs. 18/9536 v. 5.9.2016, 37; OECD (Fn. 29), Tax Risk, Tz. 11-12, 14, 112-115, 117, 145; OECD (Fn. 2), TPG, Tz. 5.10, 5.25.

[17] Vgl. Lutz / Seebeck (2019a), S. 535-543.

Grenzwerten, Vorjahreswerten oder Werten anderer konzerninterner bzw. -externer Unternehmen.[18]

Nach der CbCR-Primärpflicht (§ 138a Abs. 1 S. 1 AO) haben solche Unternehmen, die zur Aufstellung eines Konzernabschlusses verpflichtet sind, soweit ihr konsolidierter Vorjahresumsatz 750 Mio. EUR übersteigt und in ihren Konzernabschluss zumindest ein ausländisches Unternehmen oder eine ausländische Betriebsstätte einbezogen wird, ein CbCR für ihren Konzern zu erstellen.[19] Die Berichterstattung besteht dabei aus drei Tabellen. In Tabelle 1 und 2 werden alle voll- und quotenkonsolidierten Konzerngesellschaften, d.h. Kapital- und Personengesellschaften sowie Betriebsstätten, ausgewiesen.[20] Die zehn Positionen der Tabelle 1 lassen sich in Erfolgs-, Ertragsteuer- und Substanzkennzahlen unterteilen. In der Tabelle 2 sind für jede Konzerneinheit separat und unterteilt nach den Ländern ihrer Ansässigkeit bzw. Belegenheit die wichtigsten Geschäftstätigkeiten anzugeben. Sie enthält insgesamt zwölf Kategorien wichtigster Geschäftstätigkeiten und eine Auffangkategorie „Sonstige". Wird die Kategorie „Sonstige" verwendet, ist hierauf näher in Tabelle 3 einzugehen.[21] Die Tabelle 3 bietet als dritter Berichtsbestandteil schließlich Platz für weitere qualitative Angaben zu den beiden anderen Tabellen in Form von Freitextfeldern.

2.2.2.2 Technische Umsetzung & Daten

Damit die erste Risikoeinschätzung durch das CbCR ressourceneffizient erfolgen kann[22], sind berichtspflichtige Unternehmen verpflichtet, das von der OECD vorgegebene Template im digitalen Berichtsformat eXtensible Markup Language (XML) zu verwenden.[23] XML ist eine Auszeichnungssprache zur Darstellung hierarchisch strukturierter Daten.

---

[18] Vgl. OECD (2017) Tz. 37, 40, 48, 52.

[19] Siehe auch Fuchs / Steiner (2016), S. 388; Kahle / Schulz (2016), S. 819.

[20] Vgl. BT-Drs. 18/9536 v. 5.9.2016, 38; Drüen (2018), AO § 138a Rz. 14; OECD (2010) Annex III to Chapter V (S. 512). Gegen eine Einbeziehung von quotenkonsolidierten Gemeinschaftsunternehmen, vgl. Grotherr (2016), S. 995.

[21] Vgl. OECD (2010), Annex III to Chapter V (S. 511, 517).

[22] Vgl. OECD (2017), Tz. 21, 120; OECD (2010) Tz. 5.10.

[23] Vgl. BMF v. 11.7.2017 – IV B 5 - S 1300/16/10010 :002, DOK 2017/0558036, BStBl. I 2017, 974 (974).

Im Zuge des XML-basierten Reportings werden einzelne Berichtselemente, darunter sowohl Zahlen als auch Fließtexte, mit einheitlich definierten Elementen markiert – das sogenannte Tagging. Dadurch werden die CbC-Reports maschinenlesbar.

Das XML-Format schafft somit die Voraussetzungen für eine automatisierte und damit ressourcenschonende Auswertung der Reports. Gleichzeitig ist es vorteilhaft für den (internationalen) Informationsaustausch zwischen den Finanzbehörden der Länder. Es vermeidet typische Schnittstellenprobleme beim Datenaustausch, wie fehleranfällige Medienbrüche, und reduziert damit das Erfordernis manueller Anpassungen auf ein Minimum.

Computer können dank der XML-Struktur verschiedene Risikoindikatoren anhand der übermittelten Erfolgs-, Ertragsteuer- und Substanzkennzahlen aus Tabelle 1 sowie der in Tabelle 2 angezeigten Ansässigkeit bzw. Belegenheit und wichtigsten Geschäftstätigkeit für jede Konzerneinheit automatisiert errechnen. Zudem können die Informationen in den Freitextangaben aus Tabelle 3 in die automatisierte Betrachtung einbezogen werden.[24]

Die Verwendung von XML ermöglicht einen zielgerichteten Ressourceneinsatz der Finanzbehörden.[25] Sie erlaubt beispielsweise, dass die Risikoindikatoren der OECD maschinell ausgewertet werden können.

Eine derartige Vorgehensweise wird durch den Untersuchungsgrundsatz nach § 88 Abs. 5 S. 1 AO getragen, wonach automatische Risikomanagementsysteme auf Seiten der Finanzbehörde zum Einsatz kommen dürfen.

Die Möglichkeiten der computergestützten Analyse von den CbCR im XML Format sind vielfältig. Sie reichen von vergleichsweise simplen Data Mining Verfahren wie der datenbankgestützten Analyse einzelner quantitativer Werte über Benchmark- und Zeitreihenanalysen bis hin zu technisch anspruchsvollen Text Mining Verfahren wie Machine Learning-basierten Red Flag Analysen.[26]

[24] Vgl. Lutz / Seebeck (2020), S. 55.

[25] Vgl. OECD (2017) Tz. 21, 26, 120.

[26] Vgl. Seebeck / Lutz (2021), S. 260f. sowie Seebeck / Kaya (2021) und Seebeck / Vetter (2021) für mögliche Anwendungsbeispiele von Textminingverfahren im Kontext der Finanzberichterstattung.

Die OECD erachtet zu Zwecken der ersten Risikoeinschätzung zusätzliche Datenquellen als notwendig, um zutreffende Risikoeinschätzungen abgeben zu können. Diese sollen nach Möglichkeit bereits in die erste Risikoeinschätzung einfließen.[27]

2.2.2.3 Aktuelle Entwicklung: Öffentliches CbCR

Bisher haben nur die Finanzämter Zugriff auf die CbCR Daten. Die Europäische Kommission hat jedoch bereits einen Entwurf für eine Richtlinie vorgelegt, mit der Unternehmen verpflichtet werden sollen, ihre CbC-Reports zu veröffentlichen (öffentliches CbCR).[28] Hierzu soll die europäische Bilanzrichtlinie (Richtlinie 2013/34/EU) geändert werden. Die Kommission geht damit über die Arbeiten der OECD im Rahmen des BEPS-Projekts hinaus.

Dem Vorschlag der Europäischen Kommission zufolge ist das öffentliche CbCR auf der Internetseite des Mutterunternehmens fünf Jahre lang öffentlich zugänglich zu machen.[29] Die tatsächliche Zugänglichkeit (nicht der Inhalt) ist durch den Abschlussprüfer zu prüfen und entsprechend zu testieren.[30]

Kritiker argumentieren, dass ein öffentliches Country-by-Country Reporting nicht im Interesse der betroffenen Unternehmen sei, da betriebswirtschaftlich sensible Daten wie Gewinn und Umsatzerlöse öffentlich kenntlich gemacht werden würden. Konkurrenten können darüber direkte Rückschlüsse auf die Profitabilität der betroffenen Unternehmen in einzelnen Ländern ziehen, wodurch Nachteile gegenüber Konkurrenten, die nicht vom öffentlichen CbCR betroffen sind, entstehen können. Im Hinblick auf eine risikoorientierte Prüfungsplanung von Bertriebsprüfungen erscheint ein öffentliches CbCR begrüßenswert.

[27] Vgl. OECD (2017), Tz. 22, 110.

[28] Vgl. Europäische Kommission, Vorschlag für eine Richtlinie des Europäischen Parlaments und des Rates zur Änderung der Richtlinie 2013/34/EU im Hinblick auf die Offenlegung von Ertragsteuerinformationen durch bestimmte Unternehmen und Zweigniederlassungen, im Folgenden: EU-Kommission (EU) (2016) 198 v. 12.4.2016.

[29] Vgl. EU-Kommission (EU) (2016) 198, Art. 48b Abs. 1.

[30] Vgl. EU-Kommission (EU) (2016) 198, Art. 48c Abs. 3.

2.2.3 Meldepflichten steuerlicher Gestaltungen (DAC6)

2.2.3.1 Regulatorischer Hintergrund & Idee

Die Anzeigepflichten für grenzüberschreitende Steuergestaltungen der Sechsten Änderungsrichtlinie (Richtlinie (EU) 2018/822) zur EU-Amtshilferichtlinie (Richtlinie 2011/16/EU) wurden in Deutschland mit dem Gesetz zur Einführung einer Pflicht zur Mitteilung grenzüberschreitender Steuergestaltungen fristgemäß zum 1.1.2020 in nationales Recht umgesetzt. Im Folgenden werden die Sechste Änderungsrichtlinie (Richtlinie (EU) 2018/822) und die hierdurch geänderte EU-Amtshilferichtlinie (Richtlinie 2011/16/EU) zusammenfassend als DAC6-Amtshilferichtlinie (DAC6) bezeichnet (engl.: DAC6-Directive on Administrative Cooperation).

Die DAC6-Amtshilferichtlinie stellt, wie die Amtshilferichtlinie insgesamt, einen Mindeststandard dar, so dass Mitgliedstaaten darüberhinausgehende Regelungen treffen können. Das deutsche DAC6-Umsetzungsgesetz[31] setzt die Regelungen aus der DAC6-Amtshilferichtlinie unter den nach § 138 AO normierten Anzeigepflichten insbesondere in den neuen §§ 138d–138k AO sowie im Finanzverwaltungsgesetz und dem EU-Amtshilfegesetz um und entspricht weitgehend einer 1:1-Umsetzung der EU-Vorgaben.

Befeuert durch die auf den zwischenstaatlichen Steuerwettbewerb ausgerichtete Gesetzgebung einiger Staaten, wurden in den vergangenen Jahren von den Marktteilnehmern komplexe Modelle entwickelt, um die Steuerlast von individuellen Steuerpflichtigen und Unternehmen maßgeblich zu reduzieren bzw. zu vermeiden. Systematische und standardisierte Vorgehensweisen wurden entworfen, die darauf abzielen, bestehende Besteuerungslücken und -intransparenz zwischen den einzelnen Staaten zum eigenen oder im Falle der Intermediäre zum Vorteil der Kunden auszunutzen. Die Meldepflichten der DAC6 dienen dazu, derartige aggressive Steuersparmodelle zu unterbinden.[32]

Das erklärte Ziel der EU ist es, die Steuerbehörden frühzeitig über potenziell aggressive grenzüberschreitende Steuergestaltungen zu informieren.

[31] Gesetz zur Einführung einer Mitteilungspflicht für grenzüberschreitende Steuergestaltungen v. 21.12.2019, BGBl. 2019 I S. 2875.

[32] Vgl. von Brocke et al. (2021).

Dies ermöglicht etwa für die Finanzverwaltung eine frühzeitige Risikovorsorge zur zielgerichteten Betriebsprüfungstätigkeit (**Informationsfunktion**). Ein weiteres Ziel ist es, Intermediäre abzuschrecken, bei illegalen Steuerhinterziehungen und aggressiven - wenngleich legalen - Steuerumgehungen beratend tätig zu werden (**Abschreckungswirkung**).[33]

DAC6 wohnt eine doppelte Informationsfunktion inne: Erstens erhält der Gesetzgeber frühzeitig die Möglichkeit, auf aggressive Gestaltungen mit geeigneten gesetzlichen Maßnahmen zu reagieren.[34] Zweitens kann die Steuerverwaltung bei betroffenen Steuerpflichtigen gezielt ermitteln.[35] Die durch die DAC6 Meldepflichten erlangten Informationen über grenzüberschreitende Steuergestaltungen können auch zu Zwecken der Prüfungsplanung bei Betriebsprüfungen verwendet werden.

2.2.3.2 Technische Umsetzung & Daten

Die DAC6 Meldungen sind ausschließlich elektronisch, gemäß amtlich vorgeschriebenem Datensatz an das Bundeszentralamt für Steuern zu übermitteln. Dabei stehen drei Meldewege zur Verfügung:

- Einzeldatenübermittlung über das BZStOnline-Portal (BOP)
- XML-Web Upload im BOP
- Elektronische Massendatenschnittstelle (ELMA).[36]

Aufgrund der DAC6 Meldepflichten wird jährlich eine fünfstellige Anzahl an Meldungen an das BZSt erwartet.[37] Dabei ist anzunehmen, dass Intermediäre eine Vielzahl an Sachverhalten vorsichtshalber melden werden, die von den Finanzbehörden zu

[33] Vgl. Rat der EU, Verpflichtender automatischer Informationsaustausch im Bereich der Besteuerung über meldepflichtige grenzüberschreitende Modelle, Politische Einigung v. 9.3.2018 - Dok. 6804/48 FISC 103 ECOFIN 206 Nr. 2; Rat der EU, Mitteilung der Kommission über eine externe Strategie für effektive Besteuerung und Empfehlung der Kommission zur Umsetzung von Maßnahmen zur Bekämpfung des Missbrauchs von Steuerabkommen, Schlussfolgerungen des Rates v. 25.5.2016 - Dok. 9452/16 FISC 85 ECOFIN 502 Nr. 12.

[34] BMF, Schreiben zur Mitteilung grenzüberschreitender Steuergestaltungen vom 29.3.2021, V A 3 - S 0304/19/10006 :010, Tz. 3.

[35] BMF, Schreiben zur Mitteilung grenzüberschreitender Steuergestaltungen vom 29.3.2021, V A 3 - S 0304/19/10006 :010, Tz. 3

[36] Vgl. Bundeszentralamt für Steuern (2021b).

[37] Regierungsentwurf v. 9.10.2019, S. 25.

prüfen sind. Daraus resultieren erhebliche Bürokratiekosten auf der einen Seite und umfangreiche Datenbestände auf der anderen Seite. Die Kosten in den ersten drei Jahren, die auf Seiten des Bundes (BZSt, Generalzolldirektion und Informationstechnikzentrum Bund) entstehen, schätzte der Regierungsentwurf auf ca. 21-23 Mio. €.[38] Davon entfällt ein erheblicher Anteil auf den Aufbau einer leistungsfähigen IT-Infrastruktur, die u.a. auch KI-gestützte Tools zur Auswertung der Mitteilungen beinhaltet. Aus Sicht einer risikoorientierten Betriebsprüfung wäre es vielversprechend an diese IT Infrastruktur und Datenbestände anzuknüpfen.

Wie bereits im Kontext des CbCR beschrieben, ermöglicht das Übermittlungsverfahren XML ergänzend zur textuellen Beschreibung eine maschinenlesbare, logische Struktur in die Informationen zu bringen. Beispielsweise kann die Unternehmensstruktur in Form eines Baumdiagramms hierarchisch strukturiert werden. So können die Verbindung und Abhängigkeit zwischen Nutzern, verbundenen Unternehmen und betroffenen Personen einschließlich der jeweiligen Besitzverhältnisse (in Prozent) technisch eindeutig vermerkt werden. Die Erfassung der Unternehmensstruktur ist dabei jedoch auf fünf Ebenen begrenzt.[39] Derartige Informationen erscheinen im Hinblick auf die Aussteuerung von Risiken in der Betriebsprüfung nützlich.

2.2.4 Continuous Transaction Control (CTC)

2.2.4.1 Regulatorischer Hintergrund & Idee

Mit der zunehmenden Digitalisierung der staatlichen Dienste hat die International Chamber of Commerce (ICC) eine Reihe von Praxisgrundsätzen für die Implementierung kontinuierlicher Transaktionskontrollen (engl. Continuous Transaction Controls - CTCs) veröffentlicht.[40] Diese spielen eine zunehmend wichtige Rolle bei der Festlegung der Höhe der Steuerzahlungen durch die Regierungen - und können darüber hinaus zur Überwachung der Einhaltung in einer Reihe anderer Bereiche eingesetzt werden.[41]

---

[38] Regierungsentwurf v. 9.10.2019, S. 25.

[39] Vgl. BZSt, Kommunikationshandbuch Automatischer Austausch von Steuergestaltungen, Version 1.7, vom 1.04.2021, S. 28.

[40] Vgl. https://iccwbo.org/media-wall/news-speeches/continuous-transaction-controls-what-business-needs-to-know/.

[41] Vgl. https://iccwbo.org/publication/icc-continuous-transaction-control-ctcs-practice-principles/.

Im Zuge der fortschreitenden Digitalisierung haben Regierungen zunehmend cloudbasierte Dienste implementiert, um die Effizienz, Effektivität und Qualität öffentlicher Dienste zu verbessern. Diese als CTCs bezeichneten Verfahren ermöglichen es beispielsweise Strafverfolgungsbehörden aber auch den Steuerverwaltungen, Daten im Zusammenhang mit Geschäftstätigkeiten zu sammeln, die für die Ausübung ihrer Funktion relevant sind. Diese Daten werden direkt von Geschäftsdatenverwaltungssystemen in Echtzeit oder nahezu in Echtzeit abgerufen.

Dadurch ist eine Steigerung in der Effizienz der Steuererhebungsbemühungen zu erwarten. Denn CTCs beheben viele der Ineffizienzen, die mit rückwirkenden Prüfungen verbunden sind, bei denen Prüfer eine Transaktion erst lange nach ihrem Abschluss einsehen können und sich ausschließlich auf Daten stützen, die von den geprüften Unternehmen selbst gespeichert wurden. Waren die Steuerbehörden in der Vergangenheit darauf angewiesen, dass die Unternehmen historische Belege für die Hauptbücher vorlegen konnten, erfassen CTCs relevante Geschäftsinformationen beispielsweise direkt über ein Geschäftsvorgangsbuch, das aus authentifizierten Transaktionsquelldaten besteht. Auf diese Weise ermöglichen CTCs Steuerverwaltungen, Geschäftstransaktionsdaten in Echtzeit oder nahezu in Echtzeit abzurufen, wodurch die Geschwindigkeit und Genauigkeit der Steuererhebungsbemühungen verbessert werden.

2.2.4.2 Technische Umsetzung und Daten

Es werden zwei Arten von CTCs unterschieden: Berichterstattung (Reporting) und Freigabe (Clearance). Beim Reporting handelt es sich um die regelmäßige elektronische Übermittlung von Geschäftsdaten in Echtzeit an staatliche (Steuer-)Verwaltungsplattformen, ohne dass eine Genehmigung der Steuerverwaltung für diese Daten und deren fortgesetzte Verarbeitung aus steuerlicher Sicht erforderlich ist. Beim Clearance handelt es sich ebenfalls um Geschäftstransaktionsdaten, die in Echtzeit elektronisch an (Steuer-)Verwaltungsplattformen übermittelt werden. Anders als beim Reporting ist jedoch eine Genehmigung erforderlich, damit diese Daten und ihre fortgesetzte Verarbeitung aus steuerlicher Sicht gültig sind. Die Verwaltung nimmt folglich bereits im Zeitpunkt der Transaktion eine aktive Rolle ein und validiert die Daten (z.B. Rechnung), bevor die Transaktion abgeschlossen wird. Beim Reporting ist es hingegen Aufgabe der Unternehmen, die Gültigkeit der Informationen (z.B. einer Rechnung) im Nachgang der Transaktion nachzuweisen. Folglich unterscheiden sich auch die Daten je nach Modell.

Beim Reporting können grundsätzlich unterschiedliche Dateiformate zulässig sein. Während einige Länder wie Spanien und Ungarn für diese Zwecke ihren eigenen XML-Standard definiert haben, basieren andere Länder wie Portugal und Polen ihre technischen Anforderungen ganz oder teilweise auf der von der OECD herausgegebenen SAF-T-Spezifikation (Standard Audit File for Tax). Eine SAF-T-Datei enthält Rechnungsdaten sowie verschiedene andere Daten über die zugrunde liegende Lieferung aus dem ERP-System der steuerpflichtigen Unternehmen.[42]

Beim Clearance ist zwingend erforderlich, dass das berichtspflichtige Unternehmen das i.d.R. eng definierte vorgeschriebene Dateiformat verwendet, um Verzögerungen aufgrund der technischen Verarbeitung durch die Verwaltung zu verhindern.

2.2.4.3 Beispiel: My Data (Griechenland)

Mit Inkrafttreten der Richtlinie 2014/55/EU sind die zentralen öffentlichen Auftraggeber in Griechenland seit April 2019 verpflichtet, Rechnungen in einem EN-konformen Format entgegenzunehmen. Die Verpflichtung gilt seit April 2020 auch für alle anderen öffentlichen Auftraggeber. Das griechische Finanzministerium hatte 2018 zunächst mit einer „Testgruppe“ an Unternehmen begonnen, Rechnungen an die öffentliche griechische Verwaltung (B2G) und an Gesellschaften mit beschränkter Haftung in einem standardisierten Format auszustellen.[43]

Die elektronische Rechnung ist in Griechenland seit dem 1. April 2021 nunmehr auch für B2B verpflichtend. Damit erweitert das Land die bestehenden Business-to-Government (B2G) E-Invoicing-Mandate um Business-to-Business (B2B) Rechnungen. Die erste Frist für die Einführung dieser neuen Verpflichtung war ursprünglich zum Januar 2020 geplant, der Implementierungstermin wurde jedoch aufgrund der Corona-Pandemie mehrfach verschoben.

Sämtliche E-Rechnungen müssen an die MyDATA-Plattform gesendet werden.[44] Hierauf haben die griechischen Behörden uneingeschränkten Zugriff und können somit

[42] Vgl. Kowallik / Eismayr / Kirsch (2016), S. 40ff.

[43] Vgl. Horák / Bokšová / Strouhal (2020), S. 449.

[44] Vgl. Ministry of Economics, Greece (2020).

alle Daten verarbeiten und daraus anschließend Buchhaltungsdaten für jeden Steuerzahler in Griechenland erstellen. Die Übertragungsmethode für myDATA ist die myDATA REST-API.[45]

Die Rechnungen müssen das gemäß der europäischen Richtlinie vorgeschlagene elektronische Format für die Rechnungsstellung nach EN16931-1 einhalten. Derartige E-Rechnungen müssen mittels zertifizierter Software (ISO-Zertifizierung und AADE-Registrierung) ausgestellt werden.[46]

2.2.5 Aktuelle Projekte zur Aufdeckung von Steuerhinterziehung, Betrug und Geldwäsche

Neben den oben beschriebenen Entwicklungen, die den Finanzbehörden bei der Fallauswahl, Schwerpunktbildung und rationellen Prüfungsplanung und -organisation behilflich sein können, gibt es einige aktuelle Projekte und Forschungsvorhaben zur Aufdeckung von Steuerhinterziehung, Betrug und Geldwäsche o.ä., die nachfolgend kurz vorgestellt werden. Diese Projekte machen deutlich, dass weiteres Potenzial besteht, um die Betriebsprüfung durch Berücksichtigung zusätzlicher Informationsquellen und alternativer Daten effizienter zu gestalten.

2.2.5.1 Common Reporting Standard – CRS

Von der OECD wurde der „Common Reporting Standard – CRS“ zum internationalen Austausch von Finanzkonteninformationen, insbesondere Kontensalden und **bestimmten Zahlungsströmen**, entwickelt.[47] Der CRS verpflichtet Banken den Kampf gegen Steuerflucht zu unterstützen. Er hat das Ziel grenzüberschreitende Steuersachverhalte aufzudecken und dadurch Steuerhinterziehung zu erschweren. Ihm haben sich inzwischen mehr als 90 Staaten angeschlossen.[48] Zur Erhebung der Daten sind die Finanzinstitute (z.B. Banken, Versicherungen) verpflichtet.[49]

---

[45] Vgl. AADE (2020).

[46] Richtlinie 2014/55/EU des Europäischen Parlaments und des Rates vom 16. April 2014 über die elektronische Rechnungsstellung bei öffentlichen Aufträgen, Amtsblatt der Europäischen Union L 133, 6.5.2014, S. 1-11.

[47] Vgl. Bundeszentralamt für Steuern (2021a).

[48] Vgl. Bundeszentralamt für Steuern (2020).

[49] Grundlage ist das Gesetz zum automatischen Austausch von Informationen über Finanzkonten in Steuersachen (Finanzkonten-Informationsaustauschgesetz - FKAustG).

Sie melden diese in Deutschland an das Bundeszentralamt für Steuern. „Das BZSt leitet die Informationen an die zuständigen Behörden der Partnerstaaten weiter. Gleichzeitig empfängt es die Informationen aus dem Ausland und übermittelt diese dann an die zuständigen deutschen Finanzämter. Die Auswertung der Daten erfolgt in den Finanzämtern bzw. den jeweiligen Steuerbehörden im Ausland.“[50]

Die zu meldenden Finanzinformationen umfassen diverse Arten von Kapitalerträgen (unter anderem Zinsen, Dividenden, Einkünfte aus bestimmten Versicherungsverträgen und andere ähnliche Erträge), aber auch Kontoguthaben und Erlöse aus der Veräußerung von Finanzvermögen.

Die von den Finanzinstituten an das BZSt zu meldenden Daten umfassen insbesondere:

- Name, Adresse und Steueridentifikationsnummer
- Geburtsdatum und Geburtsort
- Steuerlicher Wohnsitz
- Kontonummer
- Name und Identifikationsnummer des meldenden deutschen Finanzinstituts
- Kontosaldo oder -wert zum Ende des betreffenden Kalenderjahres
- Bei Verwahrkonten jeweils der Gesamtbruttoertrag der Zinsen, der Dividenden und anderer Einkünfte, die mittels der Vermögenswerte dieses Kontos erzielt und diesem gutgeschrieben wurden
- Bei Einlagekonten der Gesamtbruttoertrag der Zinsen, die auf das Konto eingezahlt oder diesem gutgeschrieben wurden
- Bei allen anderen Konten der Gesamtbruttobetrag, der in Bezug auf das Konto an den Kontoinhaber gezahlt oder diesem gutgeschrieben wurde und für den das meldende deutsche Finanzinstitut Schuldner ist. Die Gesamthöhe aller im Meldezeitraum geleisteten Einlösungsbeträge ist einzuschließen

[50] Bundeszentralamt für Steuern (2021a).

- Bei Verwahrkonten die Gesamtbruttoerlöse aus der Veräußerung oder dem Rückkauf von Vermögensgegenständen, die auf das Konto eingezahlt oder diesem gutgeschrieben wurden und für die das Finanzinstitut als Verwahrstelle, Makler, Bevollmächtigter oder anderweitig als Vertreter für den Kontoinhaber tätig war.[51]

Damit sind vom CRS diverse Informationen betroffen, die auch aus Sicht der Betriebsprüfung von Relevanz sein können.

2.2.5.2 Ausschreibung Forschungsvorhaben des BMBF

Wie in Kapitel II.B.1 beschrieben, besteht das Ziel des CbCR darin, den Finanzbehörden zusätzliche Informationen zu grenzüberschreitenden Unternehmensgruppenstrukturen an die Hand zu geben. Die aus den von den Unternehmen gemeldeten Daten entstehende Datenbasis kann und soll für statistische Auswertungen genutzt werden, um das Ausmaß von grenzüberschreitenden Geschäftsbeziehungen innerhalb von Unternehmensgruppen zu analysieren.

Gesetzgeber und Steuerverwaltung benötigen einen wissenschaftlich fundierten und vertieften Einblick in die Datenbestände. Zu diesem Zweck wurde der Forschungsauftrag „fe 4/20: Grenzüberschreitende Geschäftsbeziehungen innerhalb von Unternehmensgruppen – Ausmaß und Reformoptionen“ ausgeschrieben, welcher das Ausmaß von grenzüberschreitenden Geschäftsbeziehungen innerhalb von Unternehmensgruppen untersuchen soll. U.a. soll im Rahmen des Forschungsprojekts untersucht werden,

- wie sich die Geschäftsaktivitäten (qualitativ und quantitativ) einer Unternehmensgruppe auf Aktivitäten zu fremden Dritten und nahestehenden Personen verteilen;
- wie hoch das Ausmaß an potenzieller Gewinnverschiebung ins (niedriger besteuernde) Ausland ist;
- wie hoch der potenzielle Verlust an deutscher Bemessungsgrundlage aufgrund von steuerlichen Gestaltungen ist.

[51] Vgl. § 8 Allgemeine Meldepflichten des Gesetzes zum automatischen Austausch von Informationen über Finanzkonten in Steuersachen (Finanzkonten-Informationsaustauschgesetz - FKAustG).

Als Datengrundlage stehen den Forschenden die länderbezogenen Berichte gemäß § 138a der Abgabenordnung zur Verfügung. Dabei sollen sowohl die aus dem Aus- als auch aus dem Inland stammenden Berichte herangezogen werden.[52]

Wie bereits oben beschrieben, sind die CbCR Daten von potenzieller Relevanz für Betriebsprüfer und somit auch die Ergebnisse des ausgeschriebenen Forschungsvorhabens.

2.2.5.3 EU-Programme in Horizont 2020 (horizon 2020)

Horizont 2020 ist das aktuelle Rahmenprogramm der Europäischen Union für Forschung und Innovation.[53] Als Förderprogramm zielt es darauf ab, EU-weit eine wissens- und innovationsgestützte Gesellschaft und eine wettbewerbsfähige Wirtschaft aufzubauen sowie gleichzeitig zu einer nachhaltigen Entwicklung beizutragen. Es gliedert sich in drei Schwerpunkte und vier zusätzliche Teilbereiche. Ein Schwerpunkt ist „Gesellschaftliche Herausforderungen".[54] Im Unterpunkt (g) „Sichere Gesellschaften – Schutz der Freiheit und Sicherheit Europas und seiner Bürger"[55] des Schwerpunkts wird unter anderem das Einzelprojekt „*Multimedia Analysis and Correlation Engine for Organised Crime Prevention and Investigation (MAGNETO)"* gefördert.[56]

MAGNETO befasst sich mit den erheblichen Bedürfnissen von Strafverfolgungsbehörden (LEAs) bei der Bekämpfung von Terrorismus und organisierter Kriminalität im Zusammenhang mit dem massiven Volumen, der Heterogenität und der Fragmentierung der Daten, die Beamte zur Verhütung, Ermittlung und Verfolgung von Straftaten analysieren müssen. Vergleichbare Herausforderungen im Umgang mit heterogenen Daten bestehen auch für die Finanzverwaltungen der Länder. Daher erscheint das Projekt auch aus Sicht einer effizienten, risikoorientierten Betriebsprüfung interessant.

---

[52] Vgl. https://www.bundesfinanzministerium.de/Content/DE/Standardartikel/Service/Ausschreibungen/2020-08-27-ausschreibung-forschungsvorhaben-fe-4-20.html (28.08.2020).

[53] https://www.horizont2020.de/einstieg-kurzueberblick.htm.

[54] Vgl. https://cordis.europa.eu/programme/id/H2020-EU.3./de.

[55] Vgl. https://cordis.europa.eu/programme/id/H2020-EU.3.7./de.

[56] Vgl. https://cordis.europa.eu/project/rcn/216142/factsheet/en.

Die Erforschung und Bereitstellung maßgeschneiderter Lösungen und Tools, die auf ausgefeilter Wissensrepräsentation, fortschrittlichem semantischem Denken und künstlicher Intelligenz basieren wie auch die Entwicklung einer internationalen gemeinsamen modularen Plattform mit integrierten Schnittstellen, welche im Rahmen der internationalen Strafverfolgung vorangetrieben wird, kann wegweisend für die digitale Betriebsprüfung sein. Ein Zugriff auf entsprechende Ressourcen und Technologien wäre für die Finanzverwaltungen zudem von großem Wert.

Mehrere massive heterogene Datenquellen zu verschmelzen und zu analysieren, verborgene Beziehungen zwischen Datenelementen aufzudecken, Trends für die Entwicklung von Sicherheitsvorfällen zu berechnen und letztendlich (und schneller) solide Beweise und beweiskräftige Unterlagen vor Gericht zu erlangen, ist eine Herausforderung, die gleichsam für Strafverfolgungsbehörden und Betriebsprüfer gilt. Auch hier lassen sich folglich Synergien heben.

2.2.6 Zwischenfazit

Es wurde deutlich, dass bereits viele Entwicklungen auf nationaler und internationaler Ebene stattfinden, die die Effizienz von klassischen, risikoorientierten Betriebsprüfungen immens steigern können. Gleichzeitig zeigen mit der Betriebsprüfung i.w.S. thematisch verwandte Projekte, dass eine enorme Ressourcenbindung und technische Ausstattung erforderlich sind, um die verfügbaren Datenmengen sinnvoll und zielgerichtet auswerten zu können.

Neben strukturellen Defiziten stellte der Bundesrechnungshof kürzlich im Hinblick auf die Umsatzsteuerbetrugsbekämpfung zahlreiche Vollzugsmängel und sogar Rückschritte bei der Betrugsbekämpfung fest.[57] Laut Bericht vom 29.10.2020 fehle ein Konzept, wie Zukunftstechnologien zur Umsatzsteuerbetrugsbekämpfung genutzt werden können. Die derzeitige IT-Unterstützung bei der Betrugsbekämpfung sei unzureichend, zentrale IT-Systeme für die umsatzsteuerlichen Kontrollen veraltet. Ferner finde aufgrund mangelnder technischer Infrastruktur kein automatisierter Datenaustausch zwischen den Zentralstellen für Betrugsbekämpfung statt.

[57] Vgl. Bundesrechnungshof (2020).

All dies sind Defizite, die es auch im Hinblick auf eine risikoorientierte, digitale Betriebsprüfung, welche alternative data zu Planungs- und Analysezwecken berücksichtigt, zu beseitigen gilt.

In anderen Ländern wie beispielsweise in Griechenland, Spanien und Italien, ist die Einführung einer elektronischen Echtzeitüberwachung (Continuous Transaction Control) in bestimmten Bereichen bereits fortgeschritten. So erhalten etwa die Finanzbehörden in Spanien oder Italien Umsatzdaten in Echtzeit und können dadurch den zeitlichen Vorsprung von aggressiv-gestaltenden oder gar betrügerischeren Steuerpflichtigen minimieren. In Griechenland wurde kürzlich die elektronische Rechnung auch für den B2B Bereich verpflichtend eingeführt. Derartige Entwicklungen sind wegweisend für eine zeitnahe, effiziente Betriebsprüfung.

Die deutsche Finanzverwaltung sollte die Chancen der Digitalisierung nutzen, um die Steuerbetrugsbekämpfung zukunftsfähig zu gestalten. Auch wenn der primäre Schwerpunkt des Umsatzsteuer-Berichts des Bundesrechnungshofs auf Verbrauchsteuern und weniger auf Gewinnverschiebungen (BEPS) liegt, sollten die Anregungen aus den Projekten zur Aufdeckung von Steuerhinterziehung, Betrug und Geldwäsche aufgegriffen werden. Ferner sollten alternative Daten verstärkt in die Prüfung einbezogen werden.

2.3 Chancen einer Nutzung von Alternativen Daten

2.3.1 Grundlagen

Der Begriff „Alternative Data“ stammt eigentlich aus der Finanzanalyse für Kapitalanlagen.

Als „Alternative Daten“ bezeichnet man hier nach verbreiteter Auffassung alle Daten, die nicht konventioneller Weise für Kapitalanlageentscheidungen genutzt werden.[58] Zu den konventionellen Daten gehören etwa die periodische Finanzberichterstattung des Unternehmens durch Jahres- und Quartalsabschlüsse, aperiodische Berichte des Unternehmens (wie Ad-hoc-Meldungen, Investorenkonferenzen, Wertpapierpros-

[58] Vgl. Monk et al. (2019).

pekte, Eintragungen ins Handelsregister), wertpapierspezifische Daten (wie Kursentwicklung und Börsenumsätze) sowie makroökonomische Rahmendaten (Zinsstrukturkurve, Inflationsraten, Wechselkurse, Bautätigkeit, Wachstumserwartungen u.a.).

Wie das Beispiel von CSR-Berichten[59] zeigt, ist dies keine statische Abgrenzung, sondern durchaus Änderungen unterworfen. Informationen über die Auswirkungen des Unternehmens auf Umwelt, Beschäftigte, Gesellschaft usw. hätte man früher noch eindeutig den Alternativen Daten zugeordnet. Im Lichte der Erkenntnis, dass CSR-Berichte u.a. auch wertvolle Hinweise über die Existenz von teilweise gravierenden Risiken liefern können, werden sie heute zunehmend auch für Kapitalanlageentscheidungen herangezogen.[60]

Als Alternative Daten in der Finanzanalyse können aktuell noch gelten:[61]

- Satellitenbilder über wirtschaftliche Aktivität (mit Autos belegte Parkplätze, Schiffsbewegungen, Landwirtschaft, Abbau von Bodenschätzen);
- Streams von Meldungen/Postings auf sozialen Medien, aus denen sich Einstellungen zu Konsum, Politik oder anderen Bereichen ableiten lassen;
- Mikrodaten über das Einkaufsverhalten von Konsumenten (z.B. Kreditkartenkäufe, In-App-Käufe bei Smartphones);
- von Webseiten im Internet extrahierte Daten (z.B. Stellenausschreibungen, um Einstellungsmuster zu erkennen);
- Datenspuren wie Chroniken, Cookies und andere digitale Fußabdrücke, die Benutzer durch ihr Surfen im Netz hinterlassen (einschließlich Geodaten von Suchen auf Smartphones).

Für Entscheidungen zur Planung von steuerlichen Betriebsprüfungen führt das Konzept der Alternativen Daten – im Sinne von konventioneller Weise nicht genutzten Daten – natürlich zu ganz anderen Ergebnissen.

---

[59] Mit der sog. CSR-Richtlinie verpflichtete die EU bestimmte große Unternehmen zur Abgabe einer „Nichtfinanziellen Erklärung“ bzw. „Konsolidierten nichtfinanziellen Erklärung“ (Richtlinie 2014/95/EU des Europäischen Parlaments und des Rates vom 22. Oktober 2014). Diese wurde in Deutschland mit den §§ 289b-289e HGB und §§ 315b-315c HGB umgesetzt.

[60] Vgl. Dhaliwal, et al. (2011); Cohen et al. (2017); Eccles et al. (2017); Amel-Zadeh / Serafeim (2018).

[61] Vgl. Monk et al. (2019).

Konventionelle Daten sind für die Betriebsprüfung zunächst einmal Daten, die der Steuerpflichtige im Rahmen seiner Steuererklärungen, Berichte, Anzeigen, Auskünfte den Finanzbehörden aktiv mitteilt oder in der Vergangenheit bereits mitgeteilt hat. Hinzu kommen Daten, welche die Finanzbehörden von anderen Steuerpflichtigen gewinnen können. Dies ermöglicht beispielsweise die Ermittlung von Richtsätzen auf statistischer Basis u.a. für verschiedene Aufwandskategorien im Verhältnis zum Umsatz oder für international übliche Lizenzraten. Ferner können Daten auch von Finanzbehörden anderer Staaten stammen. In gewissem Ausmaß lassen sich auch öffentliche Statistiken (etwa der Bundesbank) als konventionelle Daten einstufen.

Vor diesem Hintergrund ist die Bandbreite Alternativer Daten in der Betriebsprüfung potenziell enorm. Dabei handelt es sich um jedwede Daten aus Quellen außerhalb der Steuererklärung oder der sonstigen oben genannten Quellen der Finanzbehörden, die geeignet sind, die in der Steuererklärung gemachten Aussagen zu bestätigen oder eben zu widerlegen.[62] Insbesondere bieten in diesem Zusammenhang viele in der Finanzanalyse bereits verbreitet genutzte Daten ein in der Betriebsprüfung noch nicht genügend genutztes Potenzial, das es aus Sicht der Finanzbehörden zu heben gilt. So stellen beispielsweise auch Postings in sozialen Medien, online verbreitete Nachrichten oder Webseiten der Firmen zu weiten Teilen immer noch Neuland für die Betriebsprüfung dar.[63]

Die Effektivität von Betriebsprüfungen ließe sich durch die Nutzung von Alternativen Daten grundsätzlich auf zwei Ebenen steigern:

Auf **Ebene der Unternehmensauswahl** für die Betriebsprüfung werden nur die wenigsten Unternehmen zeitlich lückenlos anschlussgeprüft. In allen anderen Fällen untersucht die Betriebsprüfung die steuerlichen Verhältnisse nur für ausgewählte Jahre. Je nach Unternehmenskategorie können die zeitlichen Abstände bis zur nächsten Prüfung sehr groß sein.

[62] Die Idee externe Quellen zur risikoorientierten Prüfungsplanung einzusetzen, findet sich auch in der internen Revision. Vgl. Peemöller / Kregel (2010), S. 211.

[63] Peemöller et al. (2020), S. 310, weisen für Accounting Fraud darauf hin, dass die Aufdeckung von Bilanzdelikten in überraschend vielen Fällen nicht durch die interne Revision, Wirtschaftsprüfer oder Aufsichtsräte erfolgte, sondern (neben Whistleblowern) oft auch durch Recherchen von Journalisten, die sich öffentlich zugänglicher Quellen bedienten. Ebenso Endt et al. (2019).

Pro Jahr wird somit eine Stichprobe von Unternehmen für die Betriebsprüfung gebildet. Ziel ist es hier gezielt Unternehmen in die Stichprobe einzubeziehen, bei denen Alternative Daten Indizien liefern, die auf ein höheres Risiko an unabsichtlich oder absichtlich verkürzten Steuern hindeuten.

Auf **Ebene der Durchführung der Betriebsprüfung** können gerade bei größeren Unternehmen und Unternehmensgruppen nicht sämtliche Tatbestände überprüft werden. Hier müssen Prüfungsschwerpunkte gesetzt werden. Alternative Daten ermöglichen hier in gewisser Hinsicht eine Art der (gezielten) Stichprobenbildung. Im Sinne einer bewussten Stichprobenauswahl wären etwa zusätzliche Hinweise auf Bereiche mit erhöhtem steuerlichem Risiko sehr wertvoll. Das gilt selbst dann, wenn der Prüfer im Falle einer zeitlich lückenlosen Anschlussprüfung die Prüfungsschwerpunkte rollierend wechseln kann.

Methodisch bieten sich zwei Wege zur Ableitung von Indizien aus Alternativen Daten an:

(1) **Inkonsistenzen** zwischen den Daten gemäß den Erklärungen, Berichten, Anzeigen, Auskünften sowie weiteren Unterlagen des Steuerpflichtigen auf der einen Seite und Alternativen Daten auf der anderen Seite.

(2) Durch Alternative Daten aufgedeckte Verbindungen zu Risikobereichen gemäß einer Liste von „**Red Flags**“.[64] Diese Verbindungen können auch kettenartig über mehrere Schritte wirken.

Beide Wege sollen im Folgenden kurz umrissen werden.

2.3.2 Aufdecken von Inkonsistenzen

Von Inkonsistenz soll im Folgenden gesprochen werden, wenn zum gleichen Sachverhalt mehrere – mindestens zwei – Datenbestände vorliegen, die nicht miteinander vereinbar sind. Sie sind also widersprüchlich. Wenn wir davon ausgehen, dass die Daten einheitlich definiert wurden,[65] muss mindestens einer der Datenbestände falsch sein. Das heißt, es besteht ein Risiko falscher Steuerdaten.

[64] Ein alternativer Begriff wären schwache Signale oder „weak signals“. Vgl. Hofmann (2008), S. 412.

[65] Diese Annahme ist nicht trivial. Ein Unternehmen könnte z.B. korrekterweise gleichzeitig 300, 430 und 450 Beschäftigte haben, wenn diesen Zahlen verschiedene Definitionen zugrunde liegen: Stand

Erste Ansätze zum Aufdecken von Inkonsistenzen (Konsistenzchecks) werden bereits in Einzelfällen durch die Finanzbehörden praktiziert:

- Vergleich der in der Steuererklärung angegebenen Entfernung für den Weg zur Arbeit mit der Entfernung zwischen Adresse der Wohnung und Adresse der Arbeitsstätte gemäß Routenplanung (u.a. Google Maps, Apple Maps, Microsoft Windows Maps App).[66]
- Anbieten einer Wohnung zum Mieten auf airbnb.com mit (fehlenden) Angaben zur Vermietung in der Steuererklärung.[67]
- (Sehr) Viele Bewertungen auf Ebay.de als Folge umfangreicher Verkäufe mit (fehlenden) Angaben zu einem möglichen gewerblichen Handel in der Steuererklärung.[68]

Es ist offenkundig, dass es sich hierbei um eher „kleine Fische" handelt.

Wesentlich höheres Potenzial versprechen Ansätze, die eventuelle Steuergestaltungen größerer Unternehmen betreffen.[69] Das betrifft nicht nur große Konzerne, sondern auch viele kleine und mittlere Unternehmen.[70] Von hoher Relevanz sind insbesondere internationale Sachverhalte. Auf diese Daten hat die Finanzverwaltung keinen unmittelbaren Zugriff und die Sachverhalte betreffen nicht selten auch Jurisdiktionen, deren Behörden keine oder kaum Daten übermitteln. Außerdem kann durch gezielte Steuerplanung und -gestaltung Besteuerungssubstrat von Deutschland weg auf andere Staaten verschoben werden.

Als ein **Anwendungsbeispiel** für Konsistenzchecks kann das oben geschilderte Country-by-Country Reporting von großen Konzernen mit mindestens einer ausländischen Konzerneinheit dienen (§ 138a Abs. 1 Satz 1 AO). Diese Berichte umfassen die Listen aller in den jeweiligen Ländern ansässigen bzw. belegenen Konzerneinheiten[71]

---

am Jahresende oder im Jahresdurchschnitt? Mit oder ohne Praktikanten? Nur sozialversicherungspflichtige Arbeitnehmer oder alle? Gerechnet nach Köpfen oder Vollzeitäquivalente?

[66] Vgl. Schönwitz (2014).

[67] Vgl. Schneider (2020).

[68] Hierfür nutzt das Bundesamt für Finanzen die Software „XPider". Vgl. Schaefer-Drinhausen (2021).

[69] Zum Ausmaß u.a. Zucman (2015).

[70] Vgl. von Daniels / Wörpel (2019).

[71] CbCR-Tabelle 1. Vgl. Lutz (2020), S. 38-45.

sowie u.a. deren wichtigste Geschäftstätigkeiten[72]. Hierbei können als Freitext Erläuterungen gegeben werden.[73] Es gibt verschiedene Alternative Daten, mit denen man diese Steuerdaten manuell oder automatisiert abgleichen kann.

Eine erste Quelle bilden **Jahresabschlüsse** (Konzernabschluss und Einzel-Jahresabschlüsse). Der HGB-Einzelabschluss von Kapitalgesellschaften muss u.a. benennen: Alle Beteiligungen (§ 285 Nr. 11 HGB); alle Unternehmen, deren unbeschränkt haftender Gesellschafter die Kapitalgesellschaft ist (§ 285 Nr. 11a HGB); bei börsennotierten Kapitalgesellschaften sogar alle Anteile an großen Kapitalgesellschaften, die 5 % der Stimmrechte überschreiten. Für den HGB-Konzernabschluss wird insbesondere verlangt: Alle Tochterunternehmen (§ 313 Abs. 2 Nr. 1 HGB); alle assoziierten Unternehmen (§ 313 Abs. 2 Nr. 2 HGB); alle quotal konsolidierten Unternehmen (§ 313 Abs. 2 Nr. 3 HGB); alle Beteiligungen (§ 313 Abs. 2 Nr. 4 HGB); alle Unternehmen, deren unbeschränkt haftender Gesellschafter die Kapitalgesellschaft ist (§ 313 Abs. 2 Nr. 6 HGB); bei börsennotierten Kapitalgesellschaften ebenfalls alle Anteile an großen Kapitalgesellschaften, die 5 % der Stimmrechte überschreiten (§ 313 Abs. 2 Nr. 5 HGB). Für IFRS-Abschlüsse verlangt IFRS 12 „Angaben zu Anteilen an anderen Unternehmen“ weitreichende Hinweise.[74]

Eine zweite Quelle stellen die **Unternehmensregister** (Company Registers) der jeweiligen Länder dar. Für Deutschland ist diese das Unternehmensregister gemäß § 8b HGB (www.unternehmensregister.de), welches auch das Handelsregister mit umfasst (§ 8b Abs. 2 Nr. 1 HGB). In das Handelsregister werden bei Gründung die Namen des Eigentümers bzw. der Gesellschafter eingetragen (Einzelunternehmen § 29 HGB, offene Handelsgesellschaft § 106 Abs. 2 Nr. 1 HGB, Kommanditgesellschaft § 162 Abs. 2 HGB, GmbH § 8 Abs. 1 Nr. 3 GmbHG). Aber auch spätere Eigentümer- bzw. Gesellschafterwechsel müssen mitgeteilt werden. Durch entsprechende Aggregationen lassen sich auf dieser Basis aus dem Gesamtdatenbestand auch alle Unternehmen herausfiltern, bei denen z.B. eine bestimmte natürliche oder juristische Person Gesellschafter ist.

[72] CbCR-Tabelle 2. Vgl. Lutz (2020). S. 55-58.

[73] CbCR-Tabelle 3. Vgl. Lutz (2020). S. 58-60.

[74] Eine entsprechende Auswertung von DAX-Konzernen nimmt Redeker (2020) vor.

Bei Aktiengesellschaften ist die Situation etwas anders. Zwar ist bei der Gründung noch anzugeben, wie viele Aktien jeder Gründer übernimmt (§ 23 Abs. 2 AktG), aber das kann sich später ändern. Aktien lauten grundsätzlich auf den Namen (§ 10 Abs. 1 Satz 1 AktG). In diesem Fall ergibt sich die aktuelle Liste der Aktionäre aus dem Aktienregister der Gesellschaft gemäß § 67 AktG. Allerdings ist diese Aktionärsliste nicht öffentlich.[75] An der Stelle von Namensaktien sind unter bestimmten Bedingungen Inhaberaktien möglich, insbesondere natürlich bei Börsennotierung (§ 10 Abs. 1 Satz 2 AktG). In diesem Fall existiert kein Aktienregister, aus dem die Aktionäre hervorgehen.

Allerdings gibt es für Aktiengesellschaften eine dritte Quelle, aus der die einzelnen Aktionäre hervorgehen, zumindest wenn die Aktien an einem organisierten Markt gehandelt werden – die **Stimmrechtsmitteilungen nach WpHG**. Nach § 33 WpHG ist meldepflichtig, wer „durch Erwerb, Veräußerung oder auf sonstige Weise 3 Prozent, 5 Prozent, 10 Prozent, 15 Prozent, 20 Prozent, 25 Prozent, 30 Prozent, 50 Prozent oder 75 Prozent der Stimmrechte aus ihm gehörenden Aktien an einem Emittenten, für den die Bundesrepublik Deutschland der Herkunftsstaat ist, erreicht, überschreitet oder unterschreitet". Die Meldung muss sowohl an die Aktiengesellschaft wie an die Bundesanstalt für Finanzdienstleistungsaufsicht erfolgen. Dies gilt auch für andere, ggf. potenzielle Erwerbe von Stimmrechten z.B. über Optionen (§§ 38, 39 WpHG). Diese Stimmrechtsmitteilungen sind nach § 40 WpHG an das Unternehmensregister zu übermitteln.

Wenig beachtet wird eine weitere vierte Datenquelle. Durch die Europäische Finanzmarktverordnung[76] (**Markets in Financial Instruments Regulation**, MiFIR) haben Wertpapierfirmen, die Geschäfte mit Finanzinstrumenten tätigen, der zuständigen Behörde die vollständigen und zutreffenden Einzelheiten dieser Geschäfte zu melden (Art. 26 Abs. 1 MiFIR).

[75] Jedoch müssen in das Transparenzregister nach § 18 GwG alle sog. „wirtschaftlich Berechtigten" von juristischen Personen des Privatrechts eingetragen werden, soweit sich die geforderten Angaben (Name, Geburtsdatum, Wohnort, Art und Umfang des wirtschaftlichen Interesses; § 19 Abs. 1 GwG) nicht bereits aus einem anderen Register wie dem Unternehmensregister ergeben (§ 20 Abs. 2 GwG). Als wirtschaftlich Berechtigter gilt grundsätzlich, wer als natürliche Person direkt oder indirekt mehr als 25 % der Kapitalanteile hält, mehr als 25 % der Stimmrechte kontrolliert oder auf vergleichbare Weise Kontrolle ausübt (§ 3 GwG).

[76] Verordnung (EU) Nr. 600/2014 des Europäischen Parlaments und des Rates vom 15. Mai 2014.

Die Meldung muss Angaben zur Identifizierung der Kunden enthalten, in deren Namen die Wertpapierfirma das Geschäft abgeschlossen hat (Art. 26. Abs. 3 Satz 1 MiFIR). Art. 26 Abs. 6 MiFIR führt weiter aus: „In Bezug auf die Angaben zur Identifizierung der Kunden gemäß den Absätzen 3 und 4 verwenden Wertpapierfirmen eine Kennung für Rechtsträger (Legal Entity Identifier), die zur Identifizierung von Kunden eingeführt wurde, bei denen es sich um juristische Personen handelt." Die ESMA wurde beauftragt entsprechende technische Regulierungsstandards u.a. für Legal Entity Identifiers (LEI) zu erarbeiten (Art. 26 Abs. 9 MiFIR).

Die weltweite Vergabe von LEI-Codes wird durch die Global Legal Entity Identifier Foundation (GLEIF) geregelt, dokumentiert und überwacht.[77] Nur von der GLEIF zugelassene Vergabe-Organisationen dürfen LEIs ausgeben. Unternehmen haben die Wahl, an welche Vergabe-Organisationen sie sich zur Registrierung einer LEI wenden.[78] Der 20-stellige alphanumerische LEI-Code beruht auf der ISO-Norm 17442. Er ermöglicht eine klare und eindeutige Identifikation der Rechtsträger, die an Finanztransaktionen beteiligt sind (sog. Level 1-Daten „Wer ist wer?"). Darüber hinaus – und das ist im vorliegenden Zusammenhang von besonderer Bedeutung – sind aber für jede LEI noch Informationen über die Eigentumsstruktur des Rechtsträgers hinterlegt (sog. Level 2-Daten „Wer gehört wem?"). Alle Daten sind über die GLEIF-Webseite kostenlos öffentlich zugänglich und können auch komplett heruntergeladen werden.[79] Hieraus lässt sich (selbstverständlich nur innerhalb der erfassten Bereiche) ein weiteres, globales Verzeichnis von Unternehmensbeteiligungen ableiten.

[77] Vgl. zum Folgenden www.gleif.org/de.

[78] Eine von mehreren zertifizierten Vergabestellen in Deutschland ist etwa der Bundesanzeiger Verlag. Vgl. www.bundesanzeiger-verlag.de/evidenzwesen/leireg (3.3.2021).

[79] Vgl. www.gleif.org/de/lei-data/access-and-use-lei-data.

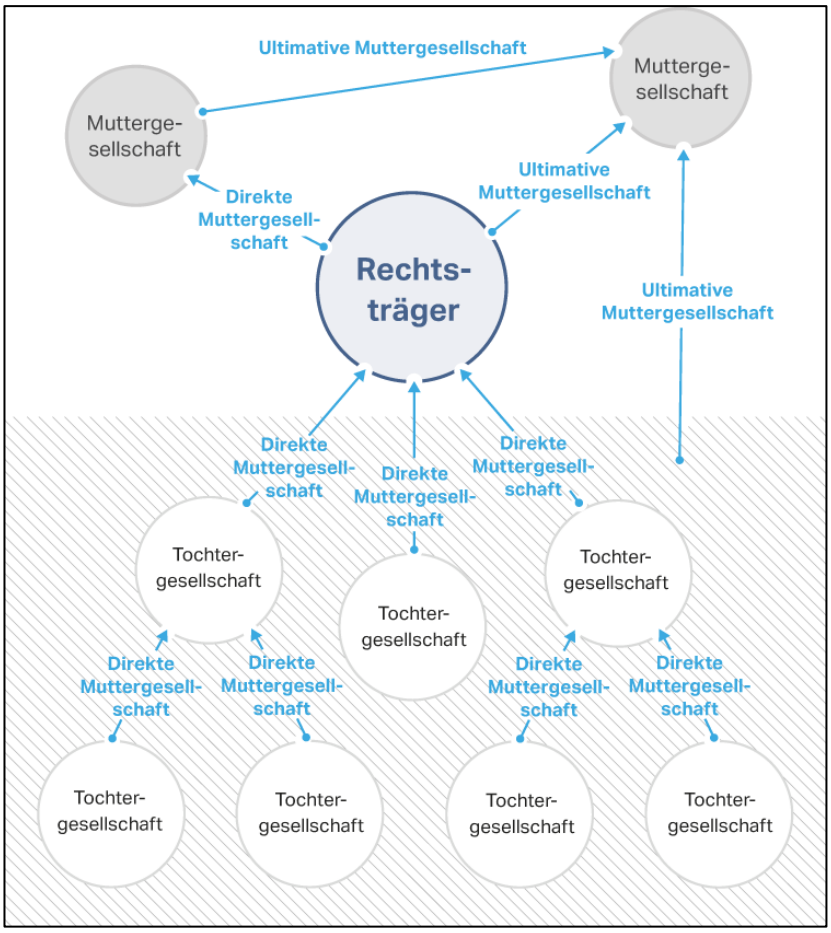


Abb. 2: Beteiligungsstrukturen aus LEI Level 2-Daten
Quelle: GLEIF, www.gleif.org/de/lei-data/access-and-use-lei-data/level-2-data-who-owns-whom (3.3.2021).

Insgesamt gibt es somit mehrere Ansatzpunkte, wie in unserem Beispiel die Liste der pro Land ansässigen bzw. belegenen Konzerneinheiten des CbCR durch Alternative Daten verprobt werden kann. Allerdings wäre eine rein manuelle Vorgehensweise in den allermeisten Fällen so aufwendig, dass dies nicht praktikabel ist.

2.3.3 Verbindungen zu Risikobereichen

2.3.3.1 Überblick

Der zweite konzeptionelle Ansatzpunkt geht von sog. „Red Flags“ aus. Hierunter versteht man in der Abschlussprüfung in der Regel Warnhinweise für Betrug (Fraud).[80]

[80] Vgl. Henselmann / Hofmann (2010), S. 268 ff.

Andere Begriffe sind „Fraud Risk Factors“ oder „Fraud Risk Indicators“. Die Prüfungsliteratur und Prüfungspraxis kennt umfangreiche Sammlungen möglicher Red Flags.[81]

Das Konzept der Red Flags lässt sich auch auf steuerliche Risiken übertragen.[82] Als Ideenspender für Risikobereiche und zugehörige Red Flags können aber nicht nur die herkömmlichen Ansatzpunkte in Betriebsprüfungen dienen. Das BEPS-Projekt der OECD greift, wie oben beschrieben, verschiedene Bereiche auf. Weitere Anhaltspunkte lassen sich aus den Projekten zur Aufdeckung von Steuerhinterziehung, Betrug und Geldwäsche ableiten. Die Grenze zwischen Verdachtspunkten für eine Betriebsprüfung, für die Steuerfahndung und/oder Geldwäsche ist in einigen Bereichen sicher fließend.

Beim Einsatz Alternativer Daten als Red Flags geht es nicht um das Erbringen eines Beweises, dass Steuern verkürzt oder gar Straftaten begangen wurden, sondern „nur“ um die Entwicklung von Indizien für Verdachtsmomente, deren Auftreten den Einsatz der Betriebsprüfung oder anderer Ermittlungsformen triggern kann. Ob es zu näheren Ermittlungen kommt, ist in der Regel eine Ermessensentscheidung der zuständigen Finanzbehörde. Falls auf der Basis von Red Flags konkrete Ermittlungen angestoßen werden, müssten diese natürlich rechtssichere Belege für tatsächlich verwirklichte – und eventuell falsch deklarierte – Sachverhalte liefern.

Henselmann / Hofmann gruppieren ihre Liste mit verschiedenen „Fraud Red Flags“ in acht sachlich verschiedene Kategorien:[83] (1) Economic or industry operating conditions; (2) Entity operating conditions (business development, organizational structure); (3) Control structure (monitoring of management, internal control components); (4) Top management; (5) Relationship between management and auditor; (6) Accounting staff; (7) Journal entries; (8) Irrational ratios.

Während die Punkte (7) *Journal entries* und (8) *Irrational ratios* bereits konventioneller Weise bei der Planung von Betriebsprüfungen genutzt werden, kann man sich in

[81] Vgl. Appendix 1: Examples of Fraud Risk Factors, International Standard on Auditing (ISA) 240, The Auditor’s Responsibilities Relating to Fraud in an Audit of Financial Statements.

[82] Vgl. Henselmann / Haller (2018).

[83] Vgl. Henselmann / Hofmann (2010), S. 268 ff.

den anderen Fällen fragen, ob eine Übertragung auf steuerliche Risiken möglich erscheint. Als Anknüpfungspunkte bieten sich an:

- die Art der Geschäftstätigkeit (analog zu (2) *Entity operating conditions – business development*);
- Standort und Rechtsform innerhalb der Konzernstruktur (analog zu (2) *Entity operating conditions – organizational structure*);
- innerhalb des Konzerns tätige Personen (analog zu (4) *Top management* sowie (6) *Accounting staff*);
- Beziehungen zu Personen außerhalb des Konzerns (analog zu (5) *Relationship between management and auditor*).

2.3.3.2 Art der Geschäftstätigkeit

Entity operating conditions deutet darauf hin, dass auch für steuerliche Risiken die Art der Geschäftstätigkeit (operations) eine Rolle spielen kann. Grundsätzlich anfällig sind **passive Betätigungen** etwa im Sinne von § 8 AStG, bei denen die Möglichkeiten zur Gewinnverschiebung besonders groß sind (z.B. Lizenzzahlungen, Verpachtung, Vermietung, Finanzierungen).

Als Red Flags können sicher auch **dubiose Betätigungen** gelten, die sich nicht selten in der Grauzone zur Legalität befinden. Beispiele hierfür sind Erotikbranche, Glücksspiel,[84] Waffenhandel, Betäubungsmittel.

Alternative Datenquellen, die Auskunft über Tätigkeiten des Unternehmens geben können, sind etwa eigene Webseiten, Nachrichten, Postings in sozialen Medien etc.

2.3.3.3 Standort und Rechtsform

Entity operating conditions umfassen zudem auch die Organisationsstruktur. Damit werden Rechtsformen und Standorte des Konzerngeflechts angesprochen. Unternehmenseinheiten (Gesellschaften und Betriebsstätten) mit Standort in **Niedrigsteuergebieten** bringen die besondere Gefahr von Gewinnverschiebungen ins Ausland mit sich. Es gibt verschiedene Listen von Niedrigsteuergebieten, sog. Steueroasen oder Tax Havens.

[84] Siehe etwa § 16 GwG.

§ 2 Abs. 2 AStG definiert „niedrige Besteuerung“ für die Wegzugsbesteuerung[85] bei natürlichen Personen als (vereinfacht ausgedrückt) Einkommensteuerbelastung, die mehr als ein Drittel geringer ist als im Inland. Dabei spielt es keine Rolle, ob die Minderung wegen einem allgemein geringeren Steuerniveau oder wegen einer eingeräumten Vorzugsbesteuerung besteht. Jedoch ist für die Besteuerung von Konzernen eher die Höhe der Körperschaftsteuer an ausländischen Standorten relevant. § 8 Abs. 3 AStG spricht bei Zwischengesellschaften von einer niedrigen Besteuerung „wenn die Einkünfte der ausländischen Gesellschaft einer Belastung durch Ertragsteuern von weniger als 25 Prozent unterliegen.“ Auch hier wäre zwischen der allgemeinen Besteuerung und Sonderregeln für bestimmte Betätigungen zu unterscheiden.

Eine Qualifikation als Steueroasen oder Tax Haven kann sich auch dadurch ergeben, dass aus manchen Ländern oder Jurisdiktionen keine oder nur **sehr eingeschränkten Informationen** über die dort ansässigen Rechtsträger und deren potentiell steuerlich relevanten Merkmale verfügbar sind.

Oxfam (2016b) unterscheidet insbesondere zwei Arten von Tax Havens:

Secrecy jurisdiction	*Corporate tax haven*
Facilitating corruption, money laundering, and avoidance and evasion of taxes on the private wealth of individuals from other countries.	Facilitating avoidance and evasion of taxes on profits of multinationals generated by operations in other countries.
No effective exchange of financial account data or ownership data.	No corporate income tax or low overall corporate tax rate.
No information available about ultimate beneficial owners	Special corporate tax regimes resulting in non-taxation of certain profits or low effective tax rates.
Legislation allowing secretive trusts and other opaque financial structures.	No effective exchange or tax rulings, country-by-country data, or other corporate tax data.
Features of corporate tax havens.	Features of secrecy jurisdictions.

Abb. 3: Tax havens – a diversified industry
Quelle: Oxfam (2016b), S. 3.

[85] Grundsätzlich beschrieben in Henselmann et al. (2005), S. 225, 502.

Aufgrund eines Beschlusses in 2016[86] führt die EU eine Liste „nicht kooperativer Länder und Gebiete“. Ziel ist es Druck auf Länder dieser Liste aufzubauen und sie zu (mehr) Kooperation zu bewegen, um Steuerbetrug, Steuervermeidung und Geldwäsche zu bekämpfen. Die aktuelle Liste vom Februar 2021 umfasst aber nur 12 Länder.[87] Insbesondere handelt es sich nur um Gebiete außerhalb der EU.

Oxfam’s Rangliste der Top 15 corporate tax havens umfasst hingegen auch Niederlande, Irland, Luxemburg sowie Zypern in der EU und daneben etwa die Schweiz und Hongkong.[88]

Auch das *Tax Justice Network,*[89] ursprünglich gegründet als Initiative im britischen Parlament,[90] veröffentlicht sowohl einen „Financial Secrecy Index“[91] wie einen „Corporate Tax Haven Index“.[92] Der Financial Secrecy Index 2020 listet auf den vorderen Plätzen nicht nur typische Kleinstaaten auf, sondern auch: 2. USA; 3. Schweiz; 4. Hongkong; 6. Luxemburg; 8. Niederlande; 12. United Kingdom. Enthalten sind ebenfalls die sog. Britischen Überseegebiete und Kronbesitzungen: 1. Cayman Islands; 9. British Virgin Islands; 11. Guernsey. Das Tax Justice Network führt dazu an: „If the UK and its network of Overseas Territories and Crown Dependencies were treated as a single entity, this UK spider’s web would rank first on the index.”[93] Verbindungen zwischen dem britischen Mutterland und Steueroasen in Britischen Überseegebieten oder Kronbesitzungen schildert sehr anschaulich als „Spinnennetz“ *Shaxson*.[94]

Auch innerhalb eines an sich unverdächtigen Staates können **bestimmte Gebiete** erhöhte Steuerrisiken mit sich bringen, wenn etwa die einzelnen Bundesstaaten oder Kantone eigene Rechtsformschriften zum Steuer- oder Gesellschaftsrecht erlassen

[86] Rat der Europäischen Union, 9452/16, FISC 85, ECOFIN 502, 25.5.2016.

[87] Rat der Europäischen Union, 2021/C 66/10, Amtsblatt der Europäischen Union C 66 vom 26.2.2021.

[88] Oxfam (2016a), S. 4.

[89] www.taxjustice.net.

[90] Vgl. Tax Justice Network (2021a).

[91] fsi.taxjustice.net/en/.

[92] www.corporatetaxhavenindex.org/en/.

[93] Tax Justice Network (2021b).

[94] Vgl. Shaxson (2012).

können. Ein bekanntes Beispiel ist der US-Bundesstaat Delaware.[95] Dort gibt es deutlich mehr Gesellschaften als Einwohner. Auf einige Einkunftsarten fallen keine Steuern an und Unternehmen können mit einem hohen Grad an Anonymität gegründet werden. Auch andere Bundesstaaten wie Nevada und Wyoming zeichnen sich entsprechend aus.[96] Hinzu kommt, dass die Vereinigten Staaten selbst den „Common Reporting Standard – CRS" der OECD zum internationalen Datenaustausch nicht unterzeichnet haben, also auch auf diesem Weg keine Informationen weitergegeben werden.[97] Ähnlich berichtete für die Schweiz die Luzerner Zeitung über auffällig viele Firmen im Verhältnis zu den Einwohnern im Kanton Zug, davon viele ohne Angestellte.[98]

Jedoch können auch **bestimmte konkrete Adressen** den Verdacht einer Briefkastenfirma nahelegen. Die schweizerische und österreichische Presse benennt Gebäude, die unglaubwürdig vielen Firmen als Anschrift dienen.

Beispielsweise beherbergt die „Bahnhofstraße 21, Zug, Schweiz" immerhin 328 aktive Firmen.[99] Aber auch In Österreich findet man im „Eckhaus Börseplatz 4/Esslinggasse 2, Wien" 124 aktive und 340 gelöschte Firmen.[100] Merkwürdig erscheint es neben der Häufung, wenn es an diesem Standort für die Firmen keinen Eintrag im Telefonverzeichnis gibt. Das spricht dann wohl wirklich für den namensgebenden Briefkasten.[101] Alternative Daten sind hier somit Unternehmensregister und Telefonverzeichnisse.

2.3.3.4 Innerhalb des Konzerns tätige Personen

Die Punkte "(4) *Top management*" und "(6) *Accounting staff*" zeigen, dass die im Konzern tätigen Personen von hoher Relevanz sein können.

---

[95] Vgl. Redeker (2020), S. 5. In dieser Studie wurde Delaware als einziger nicht eigenständiger Staat als Steueroase eingestuft.

[96] Vgl. Swanson, Ana (2016).

[97] Vgl. Bundeszentralamt für Steuern (2020).

[98] Vgl. Gwerder (2020).

[99] Vgl. Gwerder (2020).

[100] Vgl. OFR (2020).

[101] Vgl. Gwerder (2020).

Das gilt für steuerliche Risiken nicht nur in Hinblick auf eine möglicherweise **zweifelhafte Integrität** (Näheres dazu nachstehend unter e).[102]

Verschiedene Teile einer Unternehmensgruppe lassen sich nämlich nicht nur über Eigentumsrechte zusammenhalten, sondern auch über Personen (Menschen), die in Personalunion mehrere leitende oder überwachende Tätigkeiten ausüben. Nicht von ungefähr statuiert § 290 Abs. 2 Nr. 2 HGB, dass auch das Recht, „die Mehrheit der Mitglieder des die Finanz- und Geschäftspolitik bestimmenden Verwaltungs-, Leitungs- oder Aufsichtsorgans zu bestellen oder abzuberufen" einen beherrschenden Einfluss begründet. Ähnliches gilt gemäß IFRS 10.B14 (b).

Unternehmensgruppen in Form eines Konzerns sind abgestufte Gebilde. Um einen engeren Kern aus Muttergesellschaft und Tochterunternehmen (§§ 290, 294 HGB, IFRS 10) gruppieren sich weniger eng eingebundene Einheiten wie Gemeinschaftsunternehmen (§ 310 HGB, IFRS 11), Assoziierte Unternehmen (§ 311 HGB, IAS 28), Beteiligungsunternehmen (§ 271 Abs. 1 HGB) und sonstige Unternehmensanteile.

Die Konzerngrenzen lassen sich dabei nicht scharf ziehen, denn es gibt vielfältige Ermessensspielräume.[103] Solche nicht konsolidierten Gebilde können ein höheres steuerliches Risiko auslösen, wenn in der Realität der eigentlich bei Konzernfremden vorhandene Interessensgegensatz nicht existiert, da die auf den ersten Blick „fremden" Unternehmen doch in gewisser Weise „nahestehend" sind, da sie **von Personen innerhalb des Konzerns gelenkt** werden.

Für die Beurteilung von Zweckgesellschaften (§ 290 Abs. 2 Nr. 4, IFRS 10) kommt es entscheidend auf die wirtschaftliche Tragung von Chancen und Risiken an.[104] Interessant könnten Gesellschaften sein, die für den Konzernabschluss nicht als konsolidierungspflichtig eingestuft wurden. Dieser Sachverhalt wurde möglicherweise vom Konzernabschlussprüfer (§ 316 Abs. 2 HGB) untersucht und in seinem Prüfungsbericht (§ 321 HGB) erwähnt.

---

[102] Vgl. Das Framework der COSO (2013) betont als Teil des Control Environment ein „commitment to integrity and ethical values." Vgl. auch Peemöller et al. (2020), S. 345, und Hofmann (2008), S. 417.

[103] Häufig auch als Special Purpose Entities (SPE), Special Purpose Vehicle (SPV), Special Purpose Company (SPC) sowie Variable Interest Entity (VIE) bezeichnet.

[104] Vgl. Henneberger (2012).

Die Besteuerung knüpft aber grundsätzlich an den Einzelabschluss an. Selbst wenn man davon ausgeht, dass der Betriebsprüfer bei seiner Planung möglicherweise bereits Einblick in den Prüfungsbericht zum Einzelabschluss nimmt, könnte der Prüfungsbericht zum Konzernabschluss weitere Hinweise für untersuchenswerte Bereiche geben. So etwa auf nicht konsolidierte Gesellschaften, bei denen dennoch vom Konzern beschäftigte Personen als Organ oder in leitender Stellung tätig sind.

Auch außerhalb des Konzernprüfungsberichts geben Unternehmensregister Auskunft über Organmitglieder oder andere Beschäftigte des Konzerns, die zusätzlich bei nichtkonzernzugehörigen Unternehmen tätig sind.

2.3.3.5 Beziehungen zu Personen außerhalb des Konzerns

Ferner können Red Flags auch Hinweise auf eventuelle Geschäftspartner außerhalb des Konzerns geben, die für dolose Handlungen als Komplizen fungieren können. Das können natürliche oder juristische Personen sein.

Konkrete Anknüpfungspunkte für Red Flags wären Geschäftspartner Person, **über die etwas Negatives vorliegt**:

- Personen, die in der Vergangenheit wegen Vergehen (Betrug, Bestechung, Steuerhinterziehung etc.) strafrechtlich verurteilt oder zivilrechtlich belangt wurden.
- Empfänger oder Leistender von Zahlungen als Person, für die der Verdacht der Geldwäsche besteht.[105]
- Beziehungen zu Personen, die unter dem Verdacht stehen andere Straftaten zu begehen oder zu fördern.

[105] Siehe generell das Gesetz über das Aufspüren von Gewinnen aus schweren Straftaten (Geldwäschegesetz – GwG).

Solche Alternative Daten könnten u.a. aus Eintragungen ins Bundeszentralregister[106] oder ins Gewerbezentralregister,[107] von nicht-steuerlichen Behörden, aus dem Datenaustausch mit anderen Staaten oder aus bekannt gewordenen „Steuerleaks“ stammen.

Ein Verdachtsmoment sind daneben auch Beziehungen zu **intransparenten** Geschäftspartnern, insbesondere solchen, die in *secrecy jurisdictions* ansässig sind (bei Unternehmen) oder arbeiten (bei natürlichen Personen). In solchen Fällen lässt sich auch der „wirtschaftlich Berechtigte“ im Sinne des Transparenzregisters nicht ermitteln.[108] Als wirtschaftlich Berechtigte im Sinne von § 3 GwG gelten allgemein natürliche Personen, (1) in deren Eigentum oder unter deren Kontrolle der Vertragspartner letztlich steht, oder (2) auf deren Veranlassung eine Transaktion letztlich durchgeführt oder eine Geschäftsbeziehung letztlich begründet wird. Dazu zählt z.B. bei Kapitalgesellschaften, wer unmittelbar oder mittelbar mehr als 25 Prozent der Kapitalanteile hält, mehr als 25 Prozent der Stimmrechte kontrolliert oder auf vergleichbare Weise Kontrolle ausübt.

Ein gesonderter Punkt sind gegebenenfalls Beziehungen zu **politisch exponierten Personen** (z.B. aktuelle oder ehemalige Minister, hohe Verwaltungsbeamte, Richter und deren nahe Angehörige).[109] Deren Beschäftigung oder Beauftragung hat möglicherweise nicht den Grund, ihr hervorragendes Fachwissen zu nutzen, sondern dient eventuell einem (legalen) Lobbyismus[110] oder einer (illegalen) Beeinflussung ihrer

[106] Das Gesetz über das Zentralregister und das Erziehungsregister (BRZG) regelt in Deutschland die Führung dieses Registers für natürliche Personen durch das Bundesamt für Justiz (§ 1 BRZG). Neben Straftaten enthält es u.a. die Berechtigung zum Waffenbesitz und Jagderlaubnis, aber auch das Untersagen einer Berufsausübung wegen Unzuverlässigkeit, Ungeeignetheit oder Unwürdigkeit. Finanzbehörden erhalten nach aktuellem Stand (nur) Auskunft für die Verfolgung von Straftaten, die zu ihrer Zuständigkeit gehören (§ 41 Abs. 1 Nr. 4 BZRG).

[107] Nach § 149 GewO werden in das deutsche Gewerbezentralregister u.a. folgende Dinge eingetragen: Untersagung der Ausübung eines Gewerbes wegen Unzuverlässigkeit oder Ungeeignetheit, Bußgelder über 200 Euro im Zusammenhang mit der Ausübung eines Gewerbes oder wegen Steuerordnungswidrigkeiten, bestimmte Verurteilungen wegen Straftaten.

[108] Gemäß § 18 Abs. 1 GwG ist das Transparenzregister ein Register zur Erfassung und Zugänglichmachung von Angaben über den wirtschaftlich Berechtigten. Das Transparenzregister ist grundsätzlich öffentlich (§ 23 Abs. 1 Nr.3 GwG).

[109] In Deutschland definiert nach § 1 Abs. 12 und 13 GwG.

[110] Beispielsweise ist der ehemalige Bundeskanzler Gerhard Schröder u.a. Mitglied des Aufsichtsrats des russischen Ölkonzerns Rosneft, vgl. Lobbypedia (2021).

Entscheidungen zu Gunsten des Unternehmens.[111] Insofern sieht auch § 10 Abs.1 Nr. 4 GwG die Prüfung vor, ob es sich „bei dem Vertragspartner oder dem wirtschaftlich Berechtigten um eine politisch exponierte Person, um ein Familienmitglied oder um eine bekanntermaßen nahestehende Person handelt“. Hierfür gelten verstärkte Sorgfaltspflichten (§ 15 Abs. 3 Nr. 3 GwG).

2.3.4 Anwendungsbeispiel

Diese exemplarische Liste von “Red Flags“ klingt in ihren Auswirkungen erst einmal überschaubar. Das täuscht jedoch, denn ausgehend von einem einzelnen Unternehmen lassen sich Netzwerke von Verbindungen zu anderen Unternehmen oder Vermögenswerten bilden.[112]

Die konzeptionelle Idee soll an einem **einfachen fiktiven Beispiel** verdeutlicht werden:[113]

- Die XYZ-AG hat verschiedene Tochterunternehmen, u.a. die X-GmbH und die Y-GmbH.
- Die X-GmbH ist Gesellschafterin der E GmbH & Co. KG., also eines Enkelunternehmens der XYZ-AG.
- Geschäftsführerin der E GmbH & Co. KG ist Erika Musterfrau.
- Erika Musterfrau ist zugleich Allein-Geschäftsführerin der ABC-GmbH, welche nicht in den Konzernkreis der XYZ-AG einbezogen wurde.
- Die ABC-GmbH ist Lieferant der E GmbH & Co. KG.
- Die E GmbH & Co. KG und die ABC-GmbH haben außerdem die gleiche Adresse, „Hauptstraße 15, 12345 Irgendwo“.
- Die ABC-GmbH hat Zahlungen erhalten, für die der Verdacht von Geldwäsche vorliegt. Die Zahlungen liefen über eine Bank in New York und stammen von einer Person NN aus Brasilien.

[111] Peemöller et al. (2020), S. 310, weisen auf negative Wirkungen einer „Filzkultur“ in prominenten Betrugsfällen hin.

[112] Vgl. Endt et al. (2019), von Daniels / Wörpel (2019).

[113] Ein Beispiel zu Fraud Detection findet sich in Hodler / Needham (2021), S. 25ff.

- In Brasilien existiert ein Unternehmen, dessen Firmenname den Kurznamen der Konzernspitze XYZ enthält, nämlich die „XYZ Brasil Land & Building Ltd.“ (nicht in den Konzernkreis der XYZ-AG einbezogen).
- Die Y-GmbH besitzt einen 19,9 % Anteil an der Z Ltd. mit Sitz in Zypern, also unterhalb der normalen 20 % Beteiligungsschwelle.
- Mehrheitsgesellschafterin der Z-Ltd. ist die Kapitalgesellschaft P Ltd. in Panama.
- Max Mustermann als Mit-Geschäftsführer der Y-GmbH hat auf seinem Facebook-Account ein Foto in einem Restaurant in Panama City gepostet. Es zeigt ihn zusammen mit mehreren Herren im Anzug.
- Einer der Herren wird als Pedro Fulano identifiziert, ein bekannter Rechtsanwalt aus Panama.
- Pedro Fulano arbeitet in Panama in der Kanzlei „Fulano, Mengano & Zutano“, von der Steuerleaks bekannt geworden sind, die u.a. Aufschluss über höchstwahrscheinlich illegale, betrügerische Aktivitäten geben.
- Max Mustermann ist zugleich Allein-Geschäftsführer einer Kapitalgesellschaft mit Sitz in Dublin, Irland, der Atlantic Royals Ltd. Die Atlantic Royals Ltd. betreibt selbst kein operatives Geschäft. Der (einzige) Wohnort von Max Mustermann ist München.

Stellt man die oben genannten Sachverhalte bildlich dar (siehe Abb. 4), so ergibt sich durch Verkettungen untereinander ein Netz. Hierbei bilden Unternehmen, Personen, Adressen und Zahlungen die Knoten. Sie werden untereinander durch Pfeile verbunden, die Beziehungen ausdrücken. Hierzu gehören u.a. auch Eigentumsverhältnisse (besitzt_Mehrheit, besitzt_Anteil), Management-Tätigkeiten (ist_GF), Aktivitäten (reist_nach), persönliche Beziehungen (bekannt_mit) oder andere Merkmale (Name_ähnlich, hat_Tätigkeit). Dieses Netz ist mathematisch ausgedrückt ein „Graph“.

Zum Konzernkreis konsolidierter Unternehmen gehören ursprünglich die XYZ-AG (Mutter), die X-GmbH, Y-GmbH und E GmbH & Co. KG (Töchter bzw. Enkel). Sie sind in Abbildung 4 grau eingefärbt.

Im Beispiel erscheinen jedoch mehrere Verdachtsmomente naheliegend. Die ABC-GmbH residiert an derselben Adresse wie die E GmbH & Co. KG, hat dieselbe Geschäftsführerin und ist zugleich Kundin der E GmbH & Co. KG. Möglicherweise fehlt hier der Interessengegensatz und die Preise wurden entsprechend verzerrt (überhöht?) festgelegt.

Die ABC-GmbH hat aus der Lieferung Zahlungen von der E GmbH & Co. KG erhalten. Selbst leitet sie Gelder an eine Person in Brasilien weiter. Für diese Zahlungen wurden aus dem Ausland Anhaltspunkte für einen Verdacht auf Geldwäsche gemeldet. In Brasilien selbst gibt es die XYZ Brasil Land & Building Ltd., bei der eine teilweise Namensgleichheit zur XYZ-AG besteht, obwohl sie im Konzernabschluss nicht auftritt.

Die Y-GmbH besitzt Anteile an der Z Ldt. in Zypern. Mehrheitsgesellschafter der Z Ldt. ist jedoch die P Ltd. aus Panama, über die selbst nichts weiter bekannt ist. Jedoch trifft sich der Geschäftsführer der Y-GmbH in Panama anscheinend aus geschäftlichem Anlass mit dem Rechtsanwalt einer panamaischen Kanzlei, die im dringenden Verdacht krimineller Aktivitäten steht. Es könnte sich bei der P Ltd. somit auch um eine Scheinfirma handeln, hinter der in Wirklichkeit der XYZ-Konzern steht.[114]

Derselbe Geschäftsführer der Y-GmbH ist zugleich Geschäftsführer der Atlantic Royals Ltd. aus Dublin in Irland. Mit der Verwaltung von Rechten an immateriellen Wirtschaftsgütern geht sie einer passiven Tätigkeit nach. Auffällig erscheint eine geschäftsführende Tätigkeit in zwei so weit voneinander entfernten Städten, vor allem wenn es sich um Allein-Geschäftsführer handelt. An dieser Stelle ist noch offen, ob die Atlantic Royals Ltd. Geschäftsbeziehungen zum XYZ-Konzern unterhält. Dann wären gegebenenfalls die Lizenzraten zu prüfen.

[114] Zu Überschneidungen von Registerdaten mit Datenlecks aus Steueroasen vgl. Endt et al. (2019).

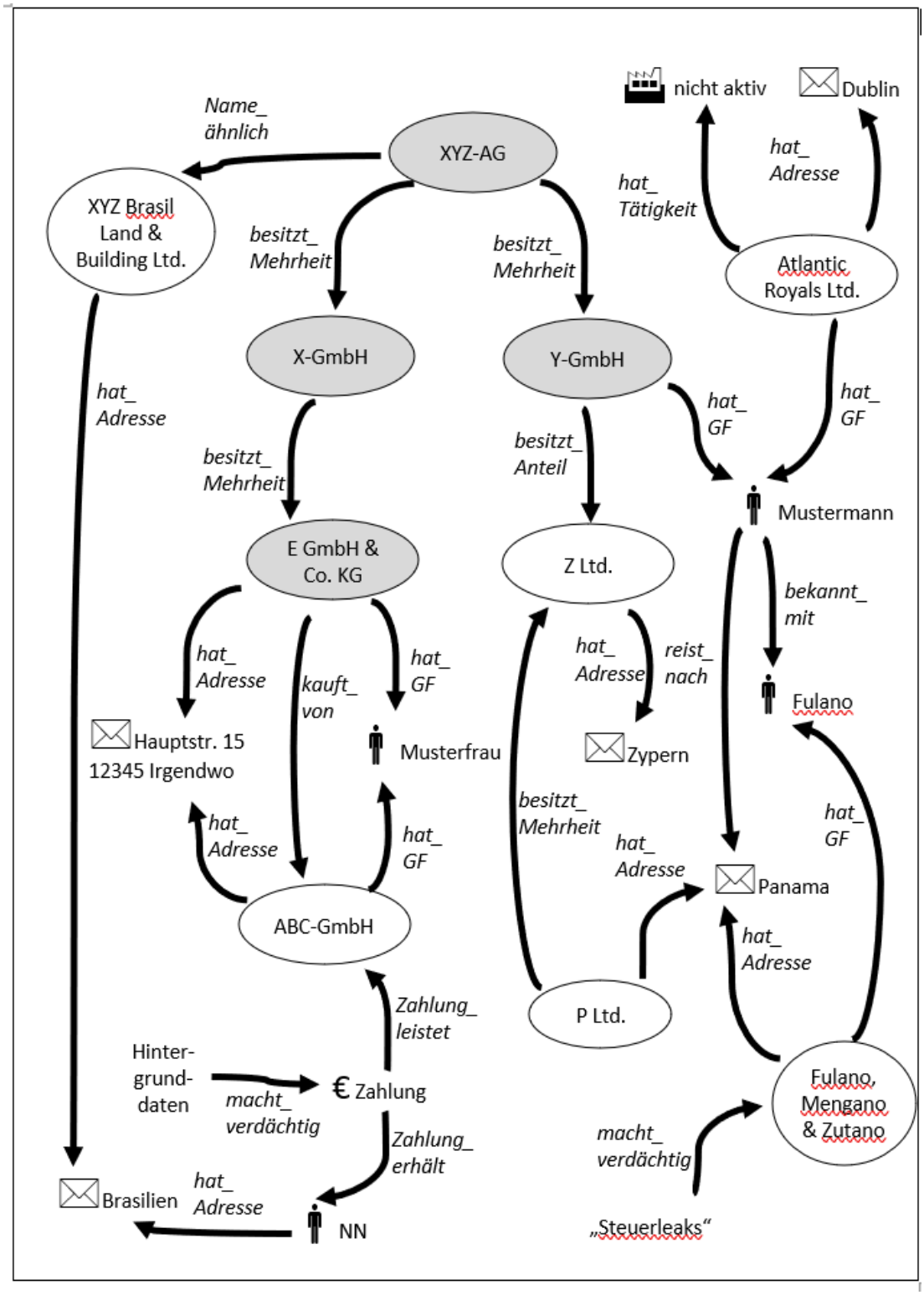


Abb. 4: Daten des Beispiels in Netzwerkdarstellung

2.3.5 Schlussfolgerungen

Die obigen Ausführungen haben Wege skizziert, wie Alternative Daten Hinweise für Betriebsprüfungen liefern können. Im Zentrum stehen als Ausgangspunkt die den Finanzbehörden bekannten Daten, insbesondere aus der Veranlagung, aber auch aus Meldepflichten für steuerliche Gestaltungen sowie aus dem Country-by-Country Reporting.

Durch stufenweise Expansion werden Verbindungen mit Alternativen Daten hergestellt, unter anderem über Unternehmensanteile, Personen, Zahlungsströme oder Adressen. Aus diesen Verbindungen entstehen entweder selbst Verdachtsmomente oder sie weisen auf Daten hin, deren Konsistenz mit den steuerlichen Unterlagen geprüft werden kann.

„Zufallsfunde“ aus Steuerleaks oder durch steuerliche „Wistleblower“ können in dieses System mit eingebaut werden (siehe Abbildung 5).[115]

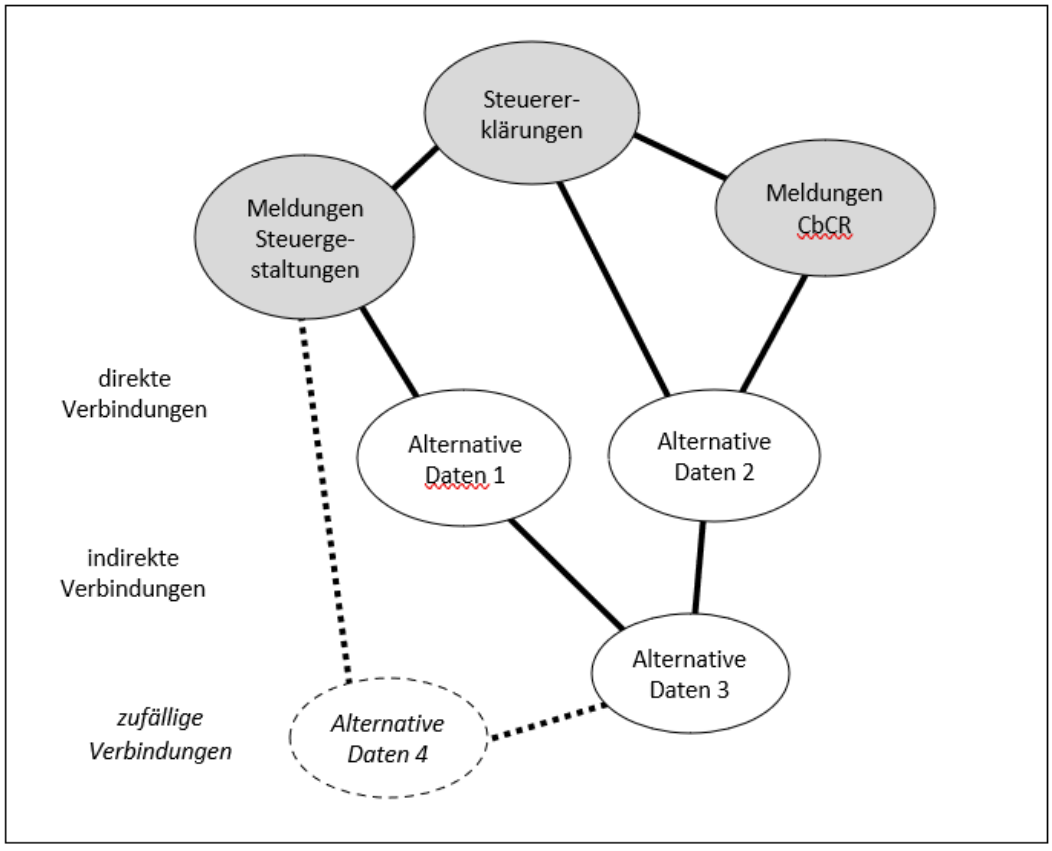


Abb. 5: Verknüpfungen der Datenbestände

[115] Vgl. von Daniels / Wörpel (2019).

Selbstverständlich bedeutet das Auftreten von „Red Flags“ nicht automatisch, dass hier wirklich eine Steuerverkürzung vorliegt.[116] Es handelt sich erst einmal nur um eine Auffälligkeit, für die zu entscheiden ist, ob man ihr weiter nachgehen will und bei der dann das Ergebnis offen ist. In vielen Fällen werden sich einzelne Anzeichen als harmlos herausstellen:

- Beispielsweise waren an einer bestimmten Adresse in der Schweiz verdächtig viele Unternehmen gemeldet. Ursache hierfür war die Tatsache, dass diese Anschrift Sitz eines Schifffahrtsunternehmens war. Für jedes der Schiffe wurde eine eigene Gesellschaft gegründet, deren Sitz dann auch an derselben Adresse lag.[117]
- Auch personelle Verbindungen können triftige außersteuerliche Gründe haben. So gibt es beispielsweise Vorstände, die bei anderen (nicht verbundenen) Konzernen Mitglied des Aufsichtsrats sind, weil dort ihre Sachkompetenz gefragt ist.

Es ist offensichtlich, dass die Chancen, eine zutreffende „Nadel im Heuhaufen“ zu finden, umso höher ist, je umfassender und umfangreicher die Alternativen Datensammlungen sind. Das steigert auch die Möglichkeiten, einzeln auftretende „Red Flags“ zu ignorieren und sich etwa auf Fälle zu konzentrieren, bei denen sich verschiedene Verdachtsmomente kumulieren.

Daher wird es kaum möglich sein die Datensammlung und Erstverarbeitung manuell vorzunehmen. Manuelle Eingriffe sollten nur nötig sein, um anhand der vom System gesammelten Anhaltspunkte diejenigen Fälle herauszugreifen, welche in der Betriebsprüfung weiterverfolgt werden sollen. Die Analyse potenzieller Alternativer Datenquellen im folgenden Kapitel beschränkt sich daher rein auf digital verfügbare Daten. Letztere ermöglichen automatisierte Verarbeitungsschritte.[118]

[116] In Hinblick auf Bilanzdelikte vgl. Peemöller et al. (2020), S. 323.

[117] Vgl. Gwerder (2020).

[118] Zu früheren Ansätzen einer wissensbasierten Fraud-Erkennung etwa durch Expertensysteme vgl. Hofmann (2008), S. 418, mit weiteren Nachweisen.

Gewisse Unschärfen in den Daten sind tragbar,[119] da wegen des bloßen Verdachtscharakters keine 100%ige Präzision und Verlässlichkeit des Datenmaterials erforderlich ist.

3 Alternative Daten

3.1 Systematisierungsansätze

In Abschnitt II. C. 1 wurden bereits Beispiele von Alternativen Daten genannt. Es gibt mehrere Möglichkeiten zur Klassifikation solcher Daten.

Nach dem **Entstehungsprozess** kann man unterscheiden, ob die Daten generiert wurden von Individuen (z.B. Posts in sozialen Netzwerken, Rezensionen auf Amazon, Suchanfragen bei Google, Downloads von SEC-Filings), durch Geschäftsprozesse (z.B. Einkauf von Waren, Käufe oder Verkäufe von Aktien, Veröffentlichung von Konzernabschlüssen, Firmengründungen, Gesellschafterwechsel) oder von Sensoren (z.B. Geo-Bewegungsprofile, Wetterdaten). Jedoch ist diese Unterscheidung für die praktische Anwendung in der Betriebsprüfung von begrenztem Wert.

Relevanter ist die Einteilung nach den **Bezugsmöglichkeiten**:

- Primärdaten werden von der Steuerverwaltung (ggf. in deren Auftrag) selbst erhoben, beispielsweise automatisiert mit Hilfe von Webscraping120 oder aber auch durch manuelle Recherchen in Registern und auf sonstigen Webseiten des Internets.
- Bei Sekundärdaten greift man auf Daten zurück, die andere bereits erhoben und i.d.R. aufbereitet haben. Dieser andere „Anbieter" kann sich auf bestimmte eng begrenzte Inhalte beschränken (sog. Point Vendor) oder aber ein Portal / ein Verzeichnis / einen Marktplatz für viele Daten unterschiedlicher Datenlieferanten bieten (sog. Platform Vendors).[121]

Wichtig sind immer die mit den Daten verbundenen **Kosten**.

[119] Beispielsweise können mehrere Personen denselben Namen haben. Umgekehrt lässt sich der Name einer Person auf mehrere Weisen darstellen.

[120] Eine gängige Python-Bibliothek ist Scrapy. Siehe https://scrapy.org/.

[121] Vgl. Monk et al. (2019). Siehe etwa https://alternativedata.org.

- Die eigene Erhebung von Daten bindet i.d.R. erhebliche Arbeitszeit. Zusätzlich ist bei maschinellen Recherchen ein gewisses technisches Know-how erforderlich, das gegebenenfalls zunächst durch Schulungen erlangt werden muss.
- Sekundärdaten von kommerziellen Anbietern müssen hingegen entgeltlich erworben werden (Private Data). Die Anforderungen an das Know-how und die benötigte Arbeitszeit sind deutlich geringer.
- Bestimmte Sekundärquellen dürfen aber auch generell oder unter bestimmten Bedingungen kostenlos genutzt werden (sog. Open Data). Dazu zählen beispielsweise Daten öffentlicher Stellen oder durch gemeinnützige Organisationen. Spezialfälle sind Open Government Data (Verwaltung) und Open Science (Wissenschaft).[122]

Damit ist bereits die Frage der **rechtlichen Nutzbarkeit** angesprochen. Für kommerzielle Anbieter richten sich die Nutzungsmöglichkeiten nach dem zugrundeliegenden Vertrag mit individuellen, maßgeschneiderten Lizenzbedingungen. Im Unterschied hierzu sind „Open Data" idealerweise Daten, die ohne Einschränkungen von jedem für jeden Zweck sowohl genutzt wie auch weiterverarbeitet und weiterverbreitet werden können.[123] Das hängt im Detail von der rechtlichen Ausgestaltung der Lizenzbedingungen ab. Am bekanntesten und verbreitetsten sind die Standardlizenzen von Creative Commons, auch CC genannt.[124] Jede Creative-Commons-Lizenz ist weltweit gültig und gilt so lange, wie der Schutz des Urheberrechts andauert.[125]

[122] Vgl. Dietrich (2011).

[123] Vgl. Dietrich (2011).

[124] Vgl. Creative Commons (2021b).

[125] Vgl. Creative Commons (2021a).

(1)	 **Attribution** **CC BY**	*This license lets others distribute, remix, adapt, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials*.
(2)	 **Attribution-ShareAlike** **CC BY-SA**	*This license lets others remix, adapt, and build upon your work even for commercial purposes, as long as they credit you and license their new creations under the identical terms. ... All new works based on yours will carry the same license, so any derivatives will also allow commercial use.*
(3)	 **Attribution-NoDerivs** **CC BY-ND**	*This license lets others reuse the work for any purpose, including commercially; however, it cannot be shared with others in adapted form, and credit must be provided to you*.
(4)	 **Attribution-NonCommercial** **CC BY-NC**	*This license lets others remix, adapt, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don't have to license their derivative works on the same terms*.
(5)	 **Attribution-NonCommercial-ShareAlike** **CC BY-NC-SA**	*This license lets others remix, adapt, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms*.
(6)	 **Attribution-NonCommercial-NoDerivs** **CC BY-NC-ND**	*This license is the most restrictive of our six main licenses, only allowing others to download your works and share them with others as long as they credit you, but they can't change them in any way or use them commercially.*

Abb. 6: Creative-Commons-Lizenzen
Eigene Abbildung - Logos und Beschreibungstexte von Creative Commons (2021a).

Für diese Varianten gilt in Hinblick auf den Einsatz zur Planung von Betriebsprüfungen:

Zu (1) Immer erforderlich ist die Namensnennung des Urhebers (Lizenzgebers), die sog. Attribution, abgekürzt „BY“. Das ist eine nur minimale Einschränkung, die lediglich die Anerkennung der Leistung des Urhebers verlangt. Diese Lizenz genügt vollständig allen Open Data Grundsätzen.

Zu (2) Eine für die Betriebsprüfung unerhebliche Einschränkung wäre, dass die Weitergabe von Daten nur unter denselben Bedingungen erfolgen darf („Share Alike“, SA).

Zu (3) Ebenfalls für die Betriebsprüfung unschädlich wäre das Verbot, die Daten in modifizierter Form zu veröffentlichen („No Derivatives“, ND), da sowieso keine Veröffentlichung beabsichtigt ist.

Zu (4) Finanzverwaltung (einschließlich der Betriebsprüfung und Steuerfahndung) sind hoheitliche, öffentlich-rechtliche Aufgaben und somit nicht-kommerzieller Natur. Daher sind auch Daten, die ausschließlich zu nicht-kommerziellen Zwecken eingesetzt werden dürfen („Non Commercial“, NC), nutzbar.

Zu (5) und (6) Gleiches gilt für die Verbindungen der Anforderungen „Non-Commercial“ mit „Share Alike“ bzw. „No Derivatives“.

Allerdings verhindern oder erschweren in der Praxis **nicht nur rechtliche Grenzen** die Nutzung von Datenbeständen. Die Open Knowledge Foundation schreibt dazu:[126] „‘*Open knowledge’ is any content, information or data that people are free to use, reuse and redistribute — without any legal, technological or social restriction. ... The key features of openness are:*

- *Availability and access: the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.*
- *Reuse and redistribution: the data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets. The data must be machine-readable.*

[126] Open Knowledge Foundation (2021).

- *Universal participation: everyone must be able to use, reuse and redistribute — there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed*."

Auch wenn Open Data grundsätzlich für den Nutzer kostenlos sein sollen, ist es möglich eine kleine Vergütung zu verlangen (siehe oben), welche die Kosten der Bereitstellung der Daten abdeckt. Das fördert die Bereitschaft, auch umfangreichere Datensammlungen zur Verfügung zu stellen. Der Nutzerkreis soll weiterhin universell sein, d.h. ein Ausschluss kommerzieller Zwecke (etwa gemäß der Creative-Commons-Lizenz „CC BY-NC") oder eine Beschränkung auf Lehr- und Forschungszwecke würde dem entgegenstehen.

Die obige Definition der Open Knowledge Foundation spricht daneben mehrere technische Aspekte an. Erstens stellt sich die Frage nach dem **Volumen**, also ob man nur einzelne konkrete Datensätze herunterladen kann oder ob ein Massendownload des Gesamtdatenbestandes möglich ist. Ersteres gilt beispielsweise für das deutsche Unternehmensregister.[127] Es enthält u.a. die Jahresabschlüsse von Kapitalgesellschaften. Jedoch muss das konkrete Unternehmen ausgewählt werden. Eine Sicherheitsabfrage nach Buchstaben/Zahlen, die auf einem Bild enthalten sind, soll den Einsatz von Webscraping-Software verhindern. Damit sind letztlich nur einzelne Abschlüsse kostenlos verfügbar. Ganz im Gegensatz hierzu steht das britische System des „Free Accounts Data Product" durch das Companies House.[128] Hier ist es möglich, als zip-Datei alle Jahresabschlüsse zusammen herunterzuladen, die in einem bestimmten Monat veröffentlicht wurden. Durch 12 Downloads erhält man so den gesamten Datenbestand des Jahres.

Zweitens verlangt die Open Knowledge Foundation eine **maschinenlesbare** Fassung der Daten. Dieses Kriterium verletzen natürlich Daten, die nur in Papierform vorliegen. Aber auch von Papierdokumenten angefertigte Scans sind zunächst nur ein Bild (Foto), ohne dass die darin enthaltenen Zeichen des Textes – Buchstaben und Zahlen – für einen Computer sofort ersichtlich sind.

[127] Vgl. www.unternehmensregister.de.

[128] Vgl. http://download.companieshouse.gov.uk/en_accountsdata.html.

Als Aufbereitungsschritt wäre zwar der Einsatz einer OCR-Software denkbar, die den gescannten Text zu erkennen sucht. Jedoch ist in der Regel keine absolut fehlerfreie Texterkennung möglich. Daher sollte es sich um eine Form von Textdateien handeln.

Die Weiterverarbeitung wird erleichtert, wenn die Daten in einer **strukturierten Form** vorliegen. Unstrukturiert wären beispielsweise die Jahresabschlüsse als pdf oder als docx (Microsoft Word), bei denen Abschlusszahlen im Text vorkommen und deren Bedeutung sich nur aus dem Kontext (Beschriftungen etc.) ergibt. Strukturiert wären hingegen Abschlussdaten in tabellarischer Form (csv oder xlsx Microsoft Excel), wenn die Spaltenköpfe und Zeilen den Inhalt eindeutig identifizieren. Weitere strukturierte Erscheinungsformen sind xml oder json-Daten, deren Bedeutung mit Hilfe von Taxonomien festgelegt werden kann sowie RDF-Daten, die auf Ontologien beruhen.[129]

Tim Berners-Lee, der Erfinder des World Wide Web, hat ein „**Fünf-Sterne-Modell**" vorgeschlagen, um den Grad der Offenheit von Daten zu kategorisieren. Die Anforderungen steigen dabei immer mehr.[130] Zusätzliche Bedingungen für höhere Stufen sind die Nutzung öffentlicher (nicht firmeneigener) Datenformate, die Nutzung von Internet-Standards zur eindeutigen Identifikation sowie zuletzt die Verlinkung mit anderen Datenbeständen:

[129] Näheres hierzu in Kapitel IV.

[130] Vgl. Dietrich (2011).

★	*Available on the web (whatever format) but with an open licence, to be Open Data*
★★	*Available as machine-readable structured data (e.g. excel instead of image scan of a table)*
★★★	*as ★★ plus non-proprietary format (e.g. CSV instead of excel)*
★★★★	*All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff*
★★★★★	*All the above, plus: Link your data to other people's data to provide context*

Abb. 7: Fünf-Sterne-Modell nach Berners-Lee
Quelle: Berners-Lee (2009)

Die folgende Tabelle fasst die wesentlichen Kriterien zusammen:

Bezugsquelle	Primärdaten der Finanzverwaltung	Sekundärdaten von anderen	
Kosten	Erhebung (z.B. Webscraping)	Kaufpreis (kommerziell)	kostenlos (Open Data)
Nutzbarkeit	voll	individuelle vertragliche Bedingungen	Lizenzbedingungen insbesondere CC
Bezugsmenge	einzelne Datensätze		Massendownload
Maschinenlesbarkeit	keine (Papier)	Bitmap-Daten (z.B. Image-Scans von Dokumenten)	Text-Daten
Strukturiertheit	unstrukturiert (z.B. Fließtext, Foto)		strukturiert (z.B. Tabellen, XML, JSON)
Formatnormierung	proprietär, durch Unternehmen (z.B. Excel)	nicht-proprietär / offen, durch Gremien	
Web-Standard	(nein)	nein (z.B. CSV, XML)	ja (durch W3C) RDF

Abb. 8: Kriterien für Alternative Daten

3.2 Nützliche Alternative Daten?

3.2.1 Vorbemerkungen

Es folgt eine Bestandsaufnahme Alternativer Datenquellen, die nach Meinung der Verfasser grundsätzlich besonders geeignet sind, Entscheidungen über eine Betriebsprüfung (Außenprüfung) – ergänzend zu den bereits intern vorliegenden steuerlichen Daten – zu unterstützen.

Dabei handelt es sich um eine subjektive Auswahl, die in keinem Falle abschließend zu verstehen ist. Die Auswahl stellt auch eine Momentaufnahme dar, da sich die Datenbestände laufend verändern und somit dynamisch sind. Ändern können sich beispielsweise Datenumfang (in der Regel Zunahme), Qualität, Aktualität, Nutzungsbedingungen oder Kosten.

Der Schwerpunkt wird auf Offene Daten (Open Data) gelegt, aber teilweise erfolgen auch Verweise auf private/kommerzielle Quellen, da hier Inhalte angeboten werden, die kaum oder nur äußerst aufwändig nachzubilden wären.

3.2.2 Unternehmen und Organe

3.2.2.1 Unternehmensregister

Ein naheliegender Startpunkt ist ein Abgleich der internen steuerlichen Daten mit den Inhalten von Unternehmensregistern.

Für **Deutschland** ist der Einstieg über die Seiten www.unternehmensregister.de oder www.bundesanzeiger.de möglich. Er beinhaltet unter einer gemeinsamen Oberfläche sowohl das Handels-, Genossenschafts- und Partnerschaftsregister (Eintragungen, eingereichte Dokumente, Registerbekanntmachungen), aber auch Rechnungslegung, Kapitalmarktinformationen im Zusammenhang mit Wertpapieren sowie insolvenzgerichtliche Bekanntmachungen.

Obwohl die Führung von Registern eine hoheitliche Aufgabe ist, wurde diese Aufgabe in Deutschland an die privatwirtschaftliche „Bundesanzeiger Verlag GmbH“ übertragen.

Diese ist aufgrund einer Privatisierungsentscheidung seit 2006 im Eigentum der Verlagsgruppe M. Dumont Schauberg.[131] Das ist u.E. eine äußerst fragwürdige Gestaltung, denn ein kostenfreier Zugriff auf die Daten wird nur im nötigen Minimum gewährt. Komfortablere und umfassendere Abfragen sind nur gegen Bezahlung möglich.

Die Bundesanzeiger Verlag GmbH schreibt dazu auf ihrer Webseite in der Rubrik „Kosten der Nutzung":[132] „*Für die Recherche nach einzelnen Firmen und die Einsicht in Veröffentlichungen und die Unternehmensträgerdaten (UT) entstehen ... keine Kosten. Für jeden Abruf der zu einer Registernummer angebotenen Daten (Aktueller Abdruck, Chronologischer Abdruck, Historischer Abdruck, Dokumentenabruf) entsteht jeweils eine Gebühr ... Die Höhe der Abrufgebühren richtet sich nach dem Justizverwaltungskostengesetz.*"

Die offiziellen Webportale der anderen **Mitgliedstaaten der Europäischen Union** finden sich auf:

https://e-justice.europa.eu/content_business_registers_in_member_states-106-en.do

Die Inhalte variieren teilweise je nach Land, aber immer sind Firma, Rechtsform, Sitz, Kapital und gesetzliche Vertreter (Organe) vorhanden. Auch für diese Staaten gilt, dass gewisse Basisinformationen kostenfrei sind und weitergehende Auskünfte in der Regel bezahlt werden müssen. Die Navigation ist natürlich immer in der Landessprache und teilweise auch in weiteren Sprachen, z.B. Englisch.

Das „**European Business Register**" (EBR)[133] ist eine freiwillige Kooperation von Unternehmensregistern in Europa, erstreckt sich aber nicht nur auf EU-Mitgliedsstaaten. Mitglieder sind u.a. auch das United Kingdom (UK), Gibraltar, die Kanalinseln Jersey und Guernsey, die Isle of Man, Liechtenstein sowie Georgien und Aserbaidschan. Die Daten stammen direkt von den teilnehmenden offiziellen nationalen Registerstellen.[134]

---

[131] https://de.wikipedia.org/wiki/Bundesanzeiger_Verlag.

[132] https://www.unternehmensregister.de/ureg/howto1.8.html. Es handelt sich um Beträge von typischerweise 4,50€ pro Unternehmen. Für alle Mitgliedstaaten der EU gilt, dass Gebühren für die Ausstellung einer vollständigen oder auszugsweisen Kopie die Verwaltungskosten nicht übersteigen dürfen (Artikel 3 Nr. 4 der Richtlinie 2009/101/EG).

[133] https://ebra.be/ (European Business Registry Association).

[134] https://ebra.be/information-distributors/.

Auf den Seiten der European Business Registry Association findet sich ebenfalls eine Übersicht von **Unternehmensregistern weiterer Staaten**.[135] Deren Inhalte richten sich nach den nationalen Vorschriften. Auch führt beispielsweise der Link für die USA auf die Webseite der „U.S. Securities and Exchange Commission“ (SEC), die (nur) über Wertpapier-emittierende Unternehmen berichtet.[136] Die eigentlichen Unternehmensregister werden dort von den Bundesstaaten geführt.

Weitere Links auf Unternehmensregister anderer Länder – ggf. auch deren innerstaatliche Teilregister nach Landesteilen – findet man auf der englischsprachigen Wikipedia-Seite „List_of_official_business_registers“.[137] Leider ist diese Seite sehr unübersichtlich und enthält auch Verweise auf Portale, die keine öffentlich-rechtlichen Unternehmensregister darstellen wie z.B. kommerzielle Datensammlungen / Auskunfteien. [138]

3.2.2.2 Identifizierung und Verknüpfung

Ein Problem der Verknüpfung von Daten aus verschiedenen Quellen besteht in der **eindeutigen Identifizierung der beteiligten Unternehmen**, etwa wenn Unternehmen A Anteile am Unternehmen B besitzt. Zu einem Unternehmen B* sind möglicherweise bereits weiterführende Informationen gespeichert. Um die Identität von B und B* zu prüfen, müsste idealerweise ein Datenquellen-übergreifender Schlüssel vorliegen.[139] Das stellt allerdings ein Problem dar. Beispielsweise ist die Adidas AG beim Amtsgericht Fürth unter der Nummer „HRB 3868“ registriert. Dieselbe Nummer „HRB 3868“ weisen aber auch u.a. die Niehaus Immobilien GmbH am Amtsgericht Freiburg oder die Herzberg Autolackiererei Verwaltungsgesellschaft mbH am Amtsgericht Kaiserslautern auf. Bereits bundesweit müsste also die Registernummer um die Angabe des Registergerichts ergänzt werden. In anderen Ländern gelten vollkommen andere Systematiken. Ideal wären weltweit gültige Schlüssel.

---

[135] https://ebra.be/worldwide-registers/.

[136] https://www.sec.gov/.

[137] https://en.wikipedia.org/wiki/List_of_official_business_registers.

[138] Vgl. Kaya / Seebeck (2019).

[139] Für natürliche Personen in Deutschland ist dies die steuerliche Identifikationsnummer (IdNr). Vgl. Bundeszentralamt für Steuern, https://www.bzst.de/DE/Privatpersonen/SteuerlicheIdentifikationsnummer/steuerlicheidentifikationsnummer_node.html.

Für dieses Problem gibt es mehrere partielle Lösungsansätze.

Als erstes ist der bereits oben in Kapitel II. C. 2. angesprochene **Legal Entity Identifier (LEI)** zu nennen.[140] Der LEI-Code ist global eindeutig. Das Datenformat wird durch die ISO-Norm 17442 definiert und verlangt strukturierte XML-Daten, welche sich sehr gut maschinell verarbeiten lassen. Unter der LEI werden u.a. folgende Informationen gespeichert:[141] offizielle Firma der juristischen Person, Adresse der Hauptniederlassung, Adresse bei Gründung, Verweis auf eine Kennung im Handelsregister (soweit vorhanden). Der LEI ist dadurch mit wesentlichen Referenzdaten verknüpft, die eine Identifikation der Rechtsträger erlauben.[142] Diese sog. „Level 1"-Daten beantworten somit die Frage „Wer ist wer?". Zusätzlich gibt es jedoch noch die „Level 2"-Daten, die Antworten auf eine für die Planung der Betriebsprüfung auch sehr wichtige Frage geben: „Wer gehört wem?" (relationship data). Unternehmen, welche einen LEI besitzen, nennen ihre „direkte buchhalterisch übergeordnete Muttergesellschaft" und ihre „ultimative buchhalterisch übergeordnete Muttergesellschaft".[143] Der komplette Datenbestand ist öffentlich zugänglich und kann tagesaktuell als Komplettdatenbestand heruntergeladen werden (ca. 300 MB).[144] Er stellt ein globales Verzeichnis von Unternehmen dar. Allerdings beschränkt sich der Kreis der Unternehmen, die eine LEI benötigen, auf juristische Personen, die Finanzinstrumente ausgeben, kaufen oder verkaufen.

Die sog. **PermID** stellt einen weiteren Ansatz für globale Unternehmensschlüssel dar.[145] Sie wird von dem Unternehmen Refinitiv – ehemals Thomson Reuters – propagiert. Refinitiv ist ein sehr großer privater Anbieter von Finanzinformationen.[146]

---

[140] Vgl. www.gleif.org.

[141] Vgl. GLEIF (2017), S. 15.

[142] Vgl. https://www.gleif.org/de/about-lei/common-data-file-format#.

[143] Vgl. https://www.gleif.org/de/about-lei/common-data-file-format/current-versions/relationship-record-cdf-format#.

[144] Download-Seite: https://www.gleif.org/de/lei-data/gleif-concatenated-file/download-the-concatenated-file.

[145] https://permid.org/.

[146] Vgl. https://www.refinitiv.com/en/about-us, https://permid.org/about.

Im Unterschied zu sonstigen Datenbeständen wird für die PermID jedoch eine kostenlose und offene Lizenz gewährt.[147] Refintiv schreibt hierzu:[148]

„*Refinitiv Permanent Identifier (PermID) is a machine readable identifier that provides a unique reference for data item. Unlike most identifiers, PermID provides comprehensive identification across a wide variety of entity types including organizations, instruments, funds, issuers and people. PermID never changes and is unambiguous, making it ideal as a reference identifier. PermID is also in the center of Refinitiv's own information model and knowledge graph.*

PermID is intended to enable interoperability. As a part of our open strategy, we provide users with a set of descriptive metadata for each entity to facilitate disambiguation to PermID with or without explicit mapping. We also provide a set of tools for working with the PermID and our Information Model, which we will continue to update, build and administer."

Die PermID ist eindeutig und die Daten liegen in einem strukturierten Format. Sie bilden auch einen Linked Data Graph. Dieser enthält neben den Unternehmensdaten im engeren Sinne („Organization") auch Daten zu Officers & Directors als natürlichen Personen. Ein Massendownload (bulk) ist möglich.[149]

[147] https://permid.org/faq: "The PermID database is licensed under the Creative Commons with Attribution license, version 4.0 (CC-BY). An extended set of fields is also available under the Creative Commons Non-Commercial license (CC-NC 3.0)."

[148] https://permid.org/.

[149] https://permid.org/faq: „Bulk download enables the retrieval of Refinitiv's entity data in .gz files, one per entity. Currently we support Turtle and Ntriples file formats."

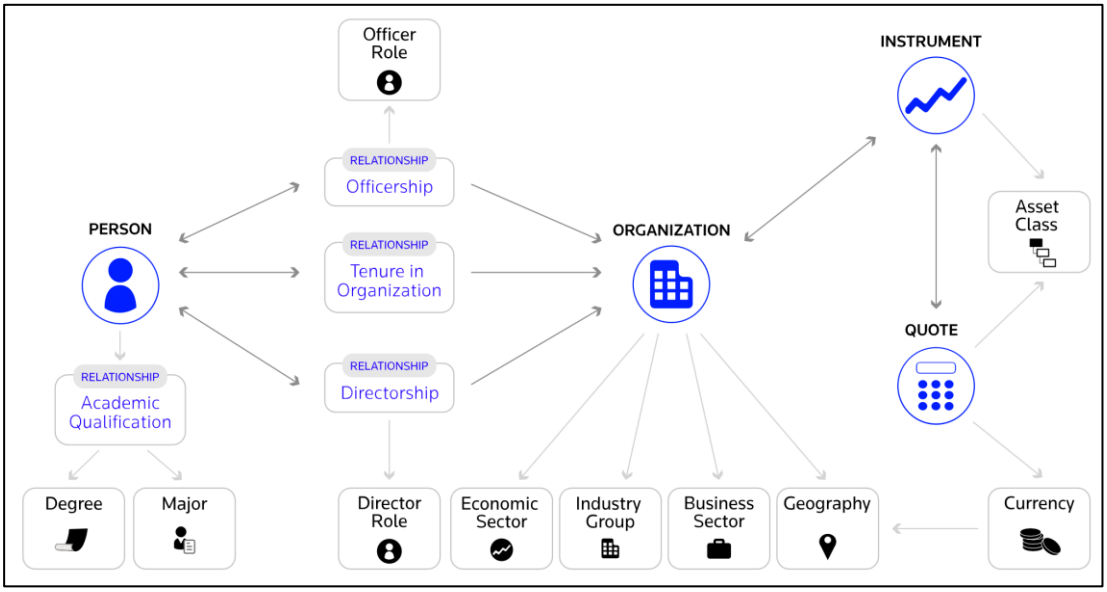


Abb. 9: Linked Data Graph
Quelle: Refinitiv, https://permid.org/faq#entitiesSupported

3.2.2.3 Aufbereitete Daten

Die oben beschriebenen Einschränkungen in der Nutzung von Registerdaten (oft kein Massendownload, originär fehlende Verbindungen zwischen den einzelnen Objekten) haben zu kommerziellen und nicht-kommerziellen Lösungen geführt.

Es gibt eine Vielzahl von privaten Unternehmen, welche **auf kommerzieller Basis** Daten von Unternehmen und deren Organen sammeln, aufbereiten und anbieten. Als Beispiel für nützliche Aufbereitungen soll die North Data GmbH (www.northdata.de) gelten:

„*North Data analysiert Pflichtveröffentlichungen europäischer Firmen und viele andere Quellen, um Wirtschaftsinformationen zu gewinnen, insbesondere zu finanziellen Kennzahlen und zu Zusammenhängen zwischen Firmen untereinander sowie zu Personen. ... Die so gewonnenen Wirtschaftsinformationen werden aufgearbeitet, um sie zu vernetzen, übersichtlich darzustellen und interaktiv zu visualisieren. Sie können sie auf dieser Website online recherchieren.*“[150]

[150] https://www.northdata.de/_about.

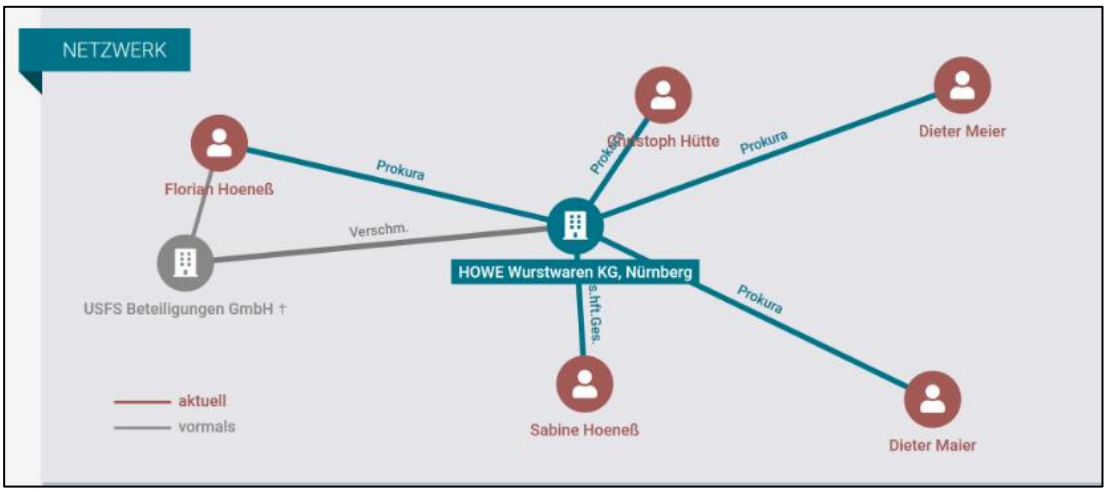


Abb. 10: Beispiel für ein Netzwerk von Unternehmen und Personen
Quelle: North Data, https://www.northdata.de

Grunddaten können kostenlos von North Data über ein Webinterface recherchiert werden. Zusätzliche Daten sowie ein Zugriff per Software über ein API sind kostenpflichtig.[151]

Im Gegensatz dazu stellt **offeneregister.de**[152] ein **nicht-kommerzielles Projekt** des „*Open Knowledge Foundation Deutschland e.V.*“[153] dar. Ausgangspunkt ist die Kritik an der Tatsache, dass die Unternehmensdaten auf der offiziellen Webseite des Handelsregisters zwar öffentlich eingesehen werden, aber nur sehr eingeschränkt genutzt werden können. Ziel ist eine Bereitstellung dieser Daten ohne Gebühren und Nutzungseinschränkungen.[154] Die Datensammlung für mehr als 5 Mio. Unternehmen beruht auf den Handelsregisterbekanntmachungen und zu einem geringen Maße den Handelsregistern selbst. Wegen rechtlicher Einschränkungen konnten die Unternehmensbekanntmachungen des Bundesanzeigers sowie Daten aus dem Unternehmensregister nicht genutzt werden.[155] Gemeinsam mit OpenCorporates.com[156] werden die Informationen als Open Data veröffentlicht. Die nachstehende Abbildung zeigt einen Vergleich mit dem offiziellen Handelsregister:

[151] Vgl. https://www.northdata.de/_premium.

[152] Vgl. https://offeneregister.de/.

[153] https://okfn.de/.

[154] Es gilt die Creative Commons Lizenz CC BY 4.0 (d.h. Attribution, also lediglich Nennung “by ...”).

[155] Vgl. https://offeneregister.de/daten/.

[156] https://opencorporates.com/.

	Handelsregister.de	Open Data
alles durchsuchen	⛔ Nein	☑ Ja, z.B. auch nach Personen
Gesamt-Download	⛔ Nein	☑ ja, z.B. hier
Freie Weiterverwendung	⛔ Nein, in Nutzungsbedingungen ausgeschlossen	☑ ja, unter freier Lizenz
Programmier-Schnittstelle	⛔ Nein	☑ vorhanden, z.B. hier oder hier
Stabile Links auf Unternehmensinfo	⛔ Nein	☑ ja, z.B. Open Knowledge Foundation Deutschland e.V.
Stabile und offene Identifikatoren	⛔ Nein	☑ wird möglich

Abb. 11: Was ermöglicht Open Data?
Quelle: Open Knowledge Foundation Deutschland e.V., https://offeneregister.de/

Damit trägt offeneregister.de auch zur **weltweiten Datenbank opencorporates.com** mit über 190 Mio. Unternehmen aus mehr als 140 Jurisdiktionen bei:[157]

„*01 Our Purpose*

Total corporate transparency is a critical requirement for a fairer society. To ensure that everyone knows exactly who they are working with – and working for. To tackle corruption and criminality. To protect our democracy. To create a trusted business environment we want to work in – and a society we'd all like to live in.

02 Our Data

As the largest open database of companies in the world, our business is making high-quality, official company data openly available. Data that can be trusted, accessed, analysed and interrogated when and how it's needed. Data that the world needs."[158]

Rechtlich gesehen wird OpenCorporates von der OpenCorporates Ltd. mit Sitz in London betrieben.[159] Jedoch werden als Unternehmenszweck (mission) Transparenz und

[157] Vgl. https://opencorporates.com/info/our-data/.

[158] https://opencorporates.com/info/about/.

[159] Company Registration Number 07444723 in England and Wales.

freier Zugang zu den Daten vorgegeben und die Einhaltung durch die Mitglieder des „OpenCorporates Trust“ überwacht.[160]

Ein Datenzugriff ist möglich über:[161]

- die Webseite (https://opencorporates.com/);
- die Programmschnittstelle OpenCorporates API (https://api.opencorporates.com/);
- als Massendownload (bulk).

Kostenlos sind generell Abfragen über die Webseite sowie für „Public Benefit Projects“.[162] Kommerzielle Nutzer können je nach Nutzungsvolumen gestaffelte API-Pläne kaufen oder ganze Massendatenbestände (bulk data).[163] Zur Identifizierung der Unternehmen werden die offiziellen Schlüssel (Identifiers) der jeweiligen Primärquellen verwendet.

3.2.3 Gesellschafter und wirtschaftlich Berechtigte

Wie bereits oben in Abschnitt II. C. 2. beschrieben, bieten sich – außerhalb der den Finanzbehörden bereits vorliegenden steuerlichen Unterlagen – folgende Wege zur Ermittlung der Gesellschafter (Miteigentümer) von Unternehmen an:

- Unternehmensregister (in Deutschland § 29 HGB, § 106 HGB, § 162 HGB, § 8 GmbHG)
- Aktienregister von Aktiengesellschaften (in Deutschland § 67 AktG)
- Anhangsangaben in Einzel- und Konzernabschlüssen (in Deutschland § 285 HGB, § 313 HGB, IFRS 12)
- Stimmrechtsmitteilungen bei Aktiengesellschaften (in Deutschland § 33 WpHG)
- LEI Level-2 Daten

[160] Vgl. https://opencorporates.com/info/governance/.

[161] Vgl. https://opencorporates.com/legal/public_records_privacy_policy.

[162] Beispielsweise investigative Journalisten, NGOs, öffentliche Forschung, möglicherweise auch staatliche Stellen zur Verfolgung von Straftaten.

[163] Vgl. https://opencorporates.com/info/our-data/.

Damit wird zwar eine Reihe von auf Eigentumsverhältnissen beruhenden Verbindungen zwischen Unternehmen sowie zwischen Privatpersonen und Unternehmen offengelegt, es verbleiben aber dennoch erhebliche Lücken. Insbesondere ist häufig nicht klar, welche natürlichen Personen letzten Endes Nutznießer der Erträge des Unternehmens sind, weil mehrere andere Organisationsstufen zwischengeschaltet sind.

Zentral für diese Frage ist der Begriff des „**wirtschaftlich Berechtigten**". Als „wirtschaftlich Berechtigter" im Sinne des Geldwäschegesetzes gilt nach § 3 Abs. 1 GwG grundsätzlich

- die natürliche Person, in deren Eigentum oder unter deren Kontrolle eine juristische Person, sonstige Gesellschaft oder eine Rechtsgestaltung wie Stiftungen, Treuhandverhältnisse o.ä. letztlich steht, oder
- die natürliche Person, auf deren Veranlassung eine Transaktion letztlich durchgeführt oder eine Geschäftsbeziehung letztlich begründet wird.

Als konkrete Fälle nennt § 3 Abs. 2-4 GwG insbesondere unmittelbare oder mittelbare Stimmrechts- oder Kapitalanteile von mehr als 25 %; Treugeber, Verwalter, Vorstände oder Begünstigte von Stiftungen o.ä.; sowie Möglichkeiten in vergleichbarer Weise Kontrolle oder beherrschenden Einfluss auszuüben.

Solche Lücken soll seit 2017 das **Transparenzregister**[164] des Geldwäschegesetzes auffüllen. Nach § 18 Abs. 2 GwG wird ein Register zur Erfassung und Zugänglichmachung von Angaben über den wirtschaftlich Berechtigten (Transparenzregister) eingerichtet. Gemäß § 26 Abs. 2 GwG wird das deutsche Transparenzregister mit den Registern anderer Mitgliedstaaten der Europäischen Union im Sinne von Artikel 22 Absatz 2 der Richtlinie (EU) 2017/1132 über die durch Artikel 22 Absatz 1 der Richtlinie (EU) 2017/1132 geschaffene zentrale Europäische Plattform vernetzt.

[164] https://www.transparenzregister.de.

Das Transparenzregister ist mittlerweile (nach Registrierung) frei öffentlich zugänglich. Dafür entstehen mit dem Abruf Gebühren.[165] Allerdings müssen zahlreiche Behörden – u.a. die Strafverfolgungsbehörden, das Bundeszentralamt für Steuern sowie die örtlichen Finanzbehörden – für ihre Einsichtnahmen nichts bezahlen.[166]

Eine wichtige Änderung erfolgte durch das am 10.6.2021 vom Bundestag verabschiedete Transparenzregister- und Finanzinformationsgesetz (TraFinG).[167] Bislang war das Transparenzregister ein sog. Auffangregister, d.h. per „Eintragungsfiktion" mussten Unternehmen keine Meldung zum Transparenzregister vornehmen, wenn die in das Transparenzregister einzutragenden Angaben zum wirtschaftlich Berechtigten schon den anderen öffentlichen Registern (z. B. Handelsregister) zu entnehmen waren. Seit dem 1.8.2021 ist das Transparenzregister jedoch ein Vollregister; die Eintragungsfiktion gilt nicht mehr.[168]

3.2.4 Unternehmensveröffentlichungen

Neben den Stammdaten wie Firma, Rechtsform, Sitz, gesetzliche Vertreter sind auch **Unternehmensveröffentlichungen in der Form von Finanzberichten** (Jahresabschlüsse, Konzernabschlüsse u.a.) von Bedeutung. Natürlich sind nicht alle Unternehmen publizitätspflichtig. Die Finanzberichte sind je nach Land bereits im Unternehmensregister enthalten oder sie müssen in einem anderen Portal aufgerufen werden.

Den Finanzberichten lassen sich potenziell allerlei nützliche Daten entnehmen:

„Allerdings gibt es in den Jahresberichten noch viel Informationen, die wir noch nicht auswerten, also zur Konzernstruktur, zum Aufsichtsrat, zur Vergütung, und vieles mehr. Das wird alles früher oder später kommen, ergänzt durch attraktive Visualisierungen."[169]

Im Falle **kapitalmarktorientierter Unternehmen** ist die Finanzberichterstattung noch deutlich umfangreicher und aussagekräftiger.

---

[165] § 24 Abs. 1, 3 GwG in Verbindung mit § 1 Transparenzregistergebührenverordnung (TrGebV) in Verbindung mit Nr. 2 Anlage 1 TrGebV.

[166] §§ 23 Abs. 1 Nr. 1, 24 Abs. 2 Satz 3 GwG.

[167] Vgl. Beck (2021).

[168] Für den Vollzug der dadurch zu ergänzenden Eintragungen gelten Übergangsfristen.

[169] Interview mit dem Gründer von Northdata, Frank Felix Debatim, in Seebach (2018).

Sie haben mehr Berichtspflichten in Form von periodischen Zwischenberichten wie Halbjahres- oder Quartalsberichten sowie einer ereignisbezogenen Berichterstattung wie Ad-hoc-Meldungen,[170] Directors‘ Dealings[171] oder Stimmrechtsmitteilungen.[172]

Diese Stimmrechtsmitteilungen können unübersichtlich werden. Allerdings führt die Bundesanstalt für Finanzdienstleistungsaufsicht eine konsolidierte Datenbank.[173] Die Angaben zu den Stimmrechtsanteilen beruhen auf den von den Emittenten veröffentlichten Stimmrechtsmitteilungen der Meldepflichtigen. Die Veröffentlichungen nach § 40 WpHG sind im elektronischen Unternehmensregister (www.unternehmensregister.de) abrufbar. Das Datum der Veröffentlichung wird in der Spalte „Veröffentlichung gemäß § 40 WpHG“ angegeben.

Vorteilhaft wirkt sich noch aus, dass auch im Ausland die Börsenzulassungsvorschriften häufig (je nach dem Kapitalmarktsegment, in dem das Unternehmen notiert ist) Berichte in englischer Sprache verlangen. Aber auch wenn dies nicht vorgeschrieben ist, würden börsennotierte Unternehmen Informationen in englischer Sprache veröffentlichen, um den Informationsbedarf internationaler Investoren zu decken. Daher unterhalten sie auf ihren Unternehmenswebseiten eine Rubrik "Investor Relations".

Berichte von kapitalmarktorientierten Unternehmen müssen auch zeitnah erstattet werden. Der Zugang zu diesen Informationen ist außerdem ohne Kosten möglich.

Auf freiwilliger Basis veröffentlichen kleine und große Unternehmen weitere Dokumente. Zu diesen gehören Pressemitteilungen, aber auch Imagebroschüren, Standorte/Niederlassungen, Interviews/Reden etc. Viele davon sind auf den Webseiten der Unternehmen verfügbar. Mit Methoden des Webscraping können solche Inhalte gesammelt werden.

---

[170] Sog. Veröffentlichung von Insiderinformationen, Art. 17 Verordnung (EU) Nr. 596/2014 (Marktmissbrauchsverordnung).

[171] Sog. Eigengeschäfte von Führungskräften, Art. 19 Verordnung (EU) Nr. 596/2014 (Marktmissbrauchsverordnung).

[172] In Deutschland nach § 33 WpHG.

[173] Vgl. https://www.bafin.de/DE/PublikationenDaten/Datenbanken/Stimmrechte/stimmrechte_node.html.

3.2.5 Steuerleaks

Steuerleaks sind „unfreiwillig“ bekannt gewordene Informationen über „graue“ oder (potentiell) illegale Aktivitäten von Unternehmen oder natürlichen Personen. Auf Wikipedia findet sich eine Liste von Leaks zu Steuerdaten.[174] Hierzu gehören[175] die sog.

- Offshore-Leaks
- Luxemburg-Leaks
- Swiss-Leaks
- Panama Papers[176]
- Bahamas-Leaks
- Malta Files
- Paradise Papers
- OpenLux
- Pandora Papers[177]

In den meisten Fällen wurde der Datensatz an das International Consortium of Investigative Journalists (ICIJ)[178] weitergegeben. Dort wurden sie in die „**ICIJ Offshore Leaks Database**“ eingespeist. Die Datenbank ist für die Öffentlichkeit verfügbar:

„*The ICIJ Offshore Leaks Database is licensed under the Open Database License and its contents under Creative Commons Attribution-ShareAlike license. Always cite the International Consortium of Investigative Journalists when using this data.*

This database is powered by Neo4j, a graph database that structures data in nodes (the icons you see in the visualization) and relationships (the links between nodes). To make this data easily accessible to everyone, regardless of the technical resources at their disposal, we have converted our original database into several CSV files, one per type of node and one for all the relationships for each project.

---

[174] https://de.wikipedia.org/wiki/Liste_von_Leaks_zu_Steuerdaten.

[175] Ohne Datensätze, die die Finanzverwaltungen aufgekauft haben und die nur innerhalb der jeweiligen Finanzverwaltung verfügbar sind.

[176] Ausführlich zur Entstehungsgeschichte vgl. Obermayer / Obermaier (2016).

[177] Ganz frisch im Herbst aufgetaucht handelt es sich um den mit Abstand größten Datenbestand im Umfang von 2,9 TB. Aktuell werden die Daten von den beteiligten Investigativ-Journalisten ausgewertet und sind noch nicht in der ICIJ verfügbar. Vgl. Balbierer et al. (2021).

[178] https://www.icij.org/.

You may download an archive in zip or torrent format. Please bear in mind that the archive is large, so if you know how to use BitTorrent, we encourage you to use it:

- *Bahamas Leaks [zip] - [torrent]*
- *Offshore Leaks [zip] - [torrent]*
- *Panama Papers [zip] - [torrent]*
- *Paradise Papers [zip] - [torrent]"*[179]

Die durchsuchbare Datenbank des International Consortium of Investigative Journalists (ICIJ) über die Panama Papers und Offshore-Leaks enthät mehr als 300.000 Einträge. Eine Suche nach Verbindungen zwischen Personen, Unternehmen und Ländern kann mühsam und enorm zeitaufwändig sein.[180]

Diese Daten wurde von der Firma Ontotext[181] – Hersteller des Graphendatenbanksystems GraphDB – genutzt und mit weiteren Open Data Quellen wie DBpedia[182] und GeoNames[183] zum „**Ontotext Linked Leaks Knowledge Graph**"[184] verbunden.

Er enthält mehr als 22 Millionen RDF Statements.[185] Auf dieser Basis können unterschiedlichste Suchen und Analysen erfolgen. Die Datenbank ist frei zugänglich auf data.ontotext.com/linkedleaks. Auch der Gesamtdatenbestand lässt sich herunterladen.[186] Das Datenbankmanagementsystem GraphDB-Free ist ebenfalls kostenfrei beziehbar.[187]

[179] https://offshoreleaks.icij.org/pages/database.

[180] Vgl. Kiryakov (2016). Die Software „Datasource" des ICIJ ist als Open Source verfügbar. Vgl. https://datashare.icij.org/.

[181] https://www.ontotext.com/.

[182] https://www.dbpedia.org/.

[183] http://www.geonames.org/.

[184] http://data.ontotext.com/linkedleaks.

[185] https://www.ontotext.com/blog/linked-leaks-a-smart-dive-into-analyzing-the-panama-papers/.

[186] Download unter ftp://ftp.ontotext.com/pub/leaks/rdf/rdf.zip.

[187] Bezug über https://www.ontotext.com/products/graphdb/graphdb-free/.

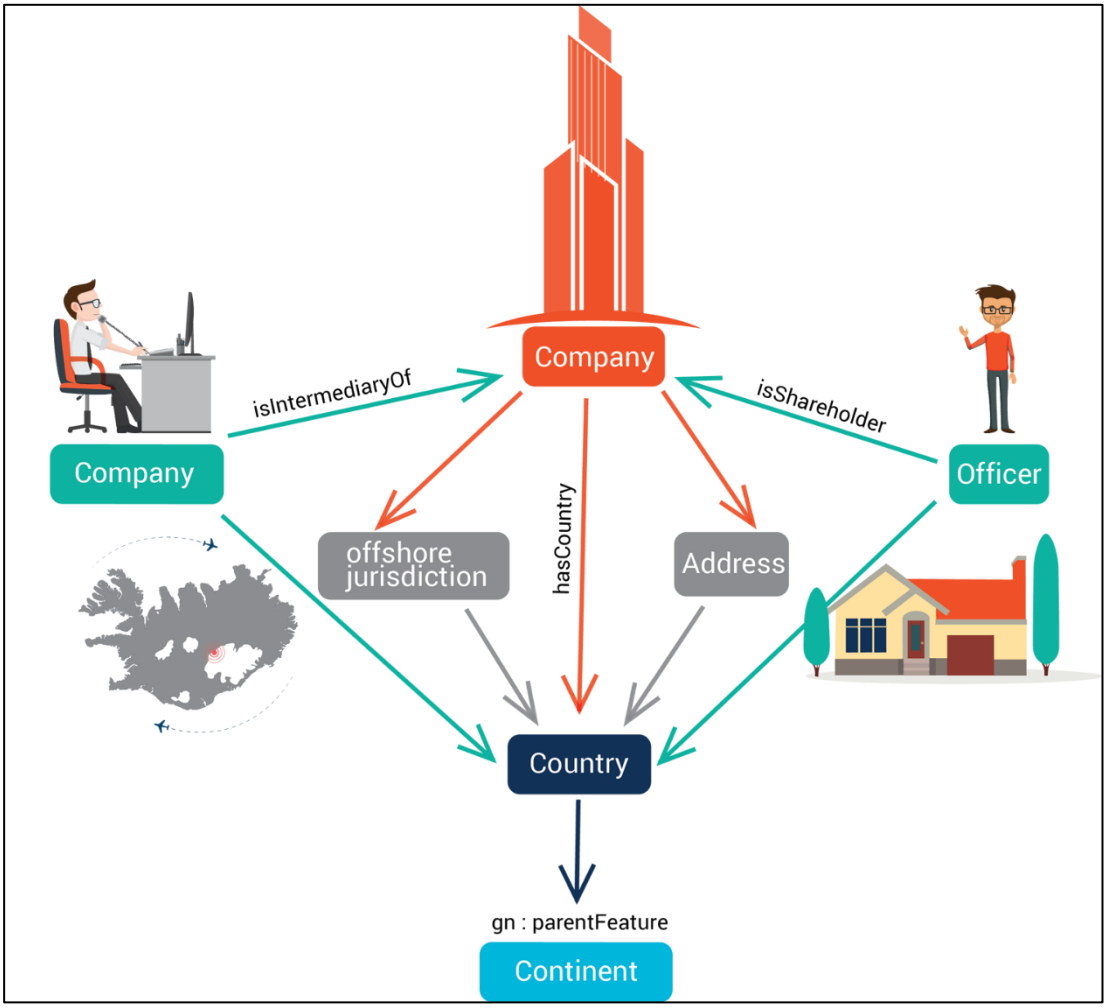


Abb. 12: 'The Game of Queries' in Linked Leaks
Quelle: Kiryakov (2016)

3.2.6 Graue Wirtschaft und Schattenwirtschaft

Bereits die Steuerleaks können dem Bereich Schattenwirtschaft zugeordnet werden. Aber auch weitere Daten geben Hinweise für wirtschaftliche Risikobereiche (siehe bereits Kapitel II. C. 3.). Hierzu gehören

- Standorte in Niedrigsteuergebieten oder Niedrigtransparenzgebieten;
- (ehemalige) Straftäter, gesuchte oder sonst wie „verdächtige" Personen als Eigentümer, Organe oder Geschäftspartner;
- auffällige Geldbewegungen und
- intrasparente Unternehmen als Geschäftspartner;

- (wegen der Möglichkeit einer illegalen Beeinflussung) Kontakte zu sog. „politisch exponierten Personen".

Umfassender als die Liste „nicht kooperativer Länder und Gebiete" der EU und als Oxfam's Rangliste der „Top 15 corporate tax havens" sind der „**Financial Secrecy Index**"[188] sowie der „**Corporate Tax Haven Index**"[189] das *Tax Justice Network*.[190]

Durch die Finanzverwaltung nutzbare Quellen stellen in Deutschland das **Bundeszentralregister**[191] und das **Gewerbezentralregister**[192] beim Bundesamt für Justiz dar (siehe auch Kapitel II. C. 3. e).

In Hinblick auf **auffällige Geldbewegungen** hat das Bundesministerium der Finanzen im Jahre 2019 die Ergebnisse der „Ersten Nationalen Risikoanalyse" veröffentlicht.[193] Als größte Risikofelder gelten anonyme Transaktionsmöglichkeiten, der Immobiliensektor, der Bankensektor (insbesondere im Rahmen des Korrespondenzbankgeschäfts und der internationalen Geldwäsche), grenzüberschreitende Aktivitäten und das Finanztransfergeschäft wegen der hohen Bargeldintensität. Das Geldwäschegesetz sieht Meldepflichten für Verpflichtete (§ 43 GwG) und Aufsichtsbehörden (§ 44 GwG) vor. Zu den verpflichteten Berufsgruppen (§ 2 GwG) zählen beispielsweise Kreditinstitute, Finanzdienstleistungsinstitute, Zahlungsinstitute, Finanzunternehmen, Versicherungsunternehmen, Versicherungsvermittler, Kapitalverwaltungsgesellschaften, Rechtsanwälte, Notare, Wirtschaftsprüfer, Steuerberater, Immobilienmakler und Glücksspielveranstalter. Die zentrale Auswertung der Verdachtsmeldungen soll durch die „Zentralstelle für Finanztransaktionsuntersuchungen" (§ 27 GwG) erfolgen. Sie ist aktuell beim Zoll – nicht den Finanzbehörden – angesiedelt und nennt sich „Financial Intelligence Unit (FIU)".[194] Allerdings gab es starke Kritik, dass die Auswertung und Weiterleitung an Strafverfolgungsbehörden nicht gut funktioniert.[195]

---

[188] fsi.taxjustice.net/en/.

[189] www.corporatetaxhavenindex.org/en/.

[190] www.taxjustice.net.

[191] https://www.bundesjustizamt.de/DE/Themen/Buergerdienste/BZR/Inhalt/Uebersicht_node.html.

[192] https://www.bundesjustizamt.de/DE/Themen/Buergerdienste/GZR/GZR_node.html.

[193] BMF (2019).

[194] https://www.zoll.de/DE/FIU/fiu_node.html.

[195] Vgl. Becker / Diehl / Traufetter (2021), Diehl (2021).

Ähnliche Daten und Analysen gibt es auch in anderen Ländern, mit denen teilweise ein Datenaustausch besteht. Nur gelegentlich werden solche Daten in der Öffentlichkeit bekannt, wie im Falle der **FinCEN-Files**. Hierbei handelt es sich um 22.000 Seiten mit Berichten über verdächtige Geldbewegungen, die von Banken weltweit an die US-Finanzaufsicht FinCEN gesendet wurden.[196] Sie wurden durch das International Consortium of Investigative Journalists (ICIJ) ausgewertet.[197] Ein Auszug von 18.153 Transaktionen ist online verfügbar.[198] Der Gesamtbestand kann aus rechtlichen Gründen nicht veröffentlicht werden, da es sich um amtliche US-Dokumente handelt.[199] Die Analysen zeigen aber, dass zahlreiche Gelegenheiten zur Austrocknung von Geldwäsche nicht genutzt wurden.[200]

Eine umfangreiche öffentliche Sammlung über Korruption und organisierte Kriminalität wird durch das **OCCRP**[201] **(Organized Crime and Corruption Reporting Project)** mit der Datensammlung **Aleph**[202] unterhalten. Der OCCRP Aleph wird durch die „Journalism Development Network, Inc." mit Sitz in Maryland, USA, einer steuerbefreiten gemeinnützigen Organisation, betrieben.[203]

Aus der Selbstdarstellung: „*As an investigative reporting platform for a worldwide network of independent media centers and journalists, OCCRP is reinventing investigative journalism as a public good. In the face of rising costs and growing threats to independent media, OCCRP provides media outlets and journalists with a range of critical resources and tools including digital and physical security and allows those covering the most sensitive topics to work in teams with trusted editors. While upholding the highest journalistic ethics and editorial standards, OCCRP develops and deploys cutting-edge tech tools to enable collaborative, secure data-driven investigations. With OCCRP Aleph, an investigative data platform powered by software we developed, journalists can search and cross-reference more than two billion records*

[196] Vgl. Engert (2020).

[197] https://www.icij.org/investigations/fincen-files/.

[198] https://www.icij.org/investigations/fincen-files/explore-the-fincen-files-data/.

[199] Vgl. Engert (2020).

[200] Vgl. Engert / Drepper (2020).

[201] https://www.occrp.org.

[202] https://aleph.occrp.org.

[203] Vgl. https://aleph.occrp.org/pages/terms.

to trace criminal connections and patterns and efficiently collaborate across borders."[204]

Die Datenbank umfasst zurzeit mehr als 300 Millionen Einträge in über 250 Datensätzen aus 140 Ländern.[205] Das gemeinsame Datenmodell des Wissensgraphen (Knowledge Graph) wird im Web beschrieben.[206]

Der Zugriff auf die Daten ist über eine Softwareschnittstelle (API) möglich, wobei die Zugriffsraten für nicht-registrierte Nutzer begrenzt sind. Für die Öffentlichkeit wird auch eine Auswahl Offener Datenbestände zum Download bereitgestellt. Angemeldete Journalisten können für ihre Projekte Massendownloads durchführen. Teil des Aleph-Software Toolkits ist Aleph Data Desktop, ein Werkzeug, mit dem man Netzwerke von Unternehmen, Personen und deren Beziehungen auf dem eigenen Computer abbilden kann. Die Software ist Open Source und wird auch von anderen Organisationen genutzt.[207]

Das kommerzielle Angebot „Watchlists & Blacklists" der info4c AG[208] aus der Schweiz sammelt internationale Warnmeldungen von Aufsichtsbehörden (Bafin, FMA, CBFA, FINMA, CNMV, MAS, FCA etc.), Fahndungsmeldungen (Interpol, FBI, DEA, DIA etc.), „disqualified directors" (Personen, denen eine Tätigkeit als Geschäftsführer untersagt ist), Sanktionslisten und manches mehr.[209]

Zu den „**politisch exponierten Personen**" (PEP) nach Art. 3 der Richtlinie (EU) 2015/849 des Europäischen Parlaments und des Rates vom 20. Mai 2015 gehören natürliche Personen, die wichtige öffentliche Ämter ausüben oder ausgeübt haben und deren Familienmitglieder sowie bekanntermaßen nahestehende Personen. Die Richtlinie wurde in Deutschland im Geldwäschegesetz umgesetzt (§ 1 Abs. 12-14 GwG). Für solche Personen gelten verschärfte Sorgfaltspflichten (§ 15 Abs. 3 Nr. 1, Abs. 4 GwG).

„Bei politisch exponierten Personen geht man aufgrund ihrer einflussreichen Position von einem höheren Risiko der Korruption und Geldwäsche aus.

---

[204] https://www.occrp.org/en/about-us.

[205] Vgl. https://aleph.occrp.org.

[206] Auf https://followthemoney.readthedocs.io/en/latest/entity.html.

[207] Vgl. die Dokumentation unter https://docs.alephdata.org/guide/getting-started.

[208] https://www.info4c.net/.

[209] Vgl. https://www.info4c.net/solutions/watchlists-blacklists/.

Nicht selten kommt es vor, dass PEPs in die Zahlung von Bestechungsgeldern verwickelt sind, um Entscheidungen oder Auftragsvergaben zu beeinflussen, Terrorismus finanzieren, Steuern hinterziehen oder illegal erworbene Gelder waschen. Die 2016 veröffentlichten Panama Papers brachten beispielsweise Daten zu 140 PEPs zum Vorschein, die Briefkastenfirmen genutzt haben sollen, um Geld zu waschen oder um sich selbst als Eigentümer unsauberen Geldes zu vertuschen.“[210]

Listen politisch exponierter Personen (PEP-Listen) werden von verschiedenen kommerziellen Datenlieferanten geführt, beispielsweise „Nexis Diligence“ von LexisNexis[211] oder „PEP Desk“ von info4c.[212]

Es gibt jedoch auch frei zugängliche Datensammlungen, etwas durch „opensanctions.org“.[213] Hier zu finden ist u.a. eine Aggregation „Politically Exposed Persons (PEPs)“ aus 5 verschiedenen Datenquellen mit insgesamt über 87.000 Einträgen.[214] Darunter sind beispielsweise auch frühere oder ehemalige Mitglieder des Deutschen Bundestages oder die Mitglieder des Europäischen Parlaments zu finden.

3.2.7 Nachrichten und Ereignisse

Ein Vergleich der den Finanzbehörden vorliegenden Daten mit Nachrichten in der Presse oder andere Meldungen über Ereignisse kann ebenfalls Unstimmigkeiten zu Tage treten lassen oder bisher unentdeckte Verbindungen aufdecken.

Aufgrund der Vielzahl von (Presse-)Medien wird es in der Regel keinen Sinn machen, die einzelnen Medienquellen individuell auszuwerten. Sinnvoller ist ein Zugriff auf Aggregatoren, die eine Vielzahl von Medien erfassen. Beispielhaft seien genannt:

- **Google News**.[215] Die (Links auf) die Original-Nachrichten werden von einer Vielzahl von Quellen zusammengetragen. Es gibt keine Aufbereitung, aber die Möglichkeit einer Personalisierung, d.h. die Wahl eigener Quellen und Themen.

[210] https://www.lexisnexis.de/begriffserklaerungen/compliance/pep-politisch-exponierte-personen.

[211] Vgl. https://www.lexisnexis.de/loesungen/compliance/geschaeftspartnerpruefung-diligence.

[212] Vgl. https://www.info4c.net/solutions/pep-desk-database/.

[213] Vgl. https://opensanctions.og.

[214] Vgl. https://opensanctions.org/datasets/peps/.

[215] https://news.google.com.

- **Global Entity Graph** (GEG) des "**GDELT Project**" (Global Database of Events, Language, and Tone).[216] "*GDELT monitors print, broadcast, and web news media in over 100 languages from across every country in the world to keep continually updated on breaking developments anywhere on the planet. Its historical archives stretch back to January 1, 1979 and update every 15 minutes. ... The GDELT Translingual platform represents what we believe is the largest realtime streaming news machine translation deployment in the world: all global news that GDELT monitors in 65 languages, representing 98,4 % of its daily non-English monitoring volume, is translated in realtime into English and processed.*"[217] Es identifiziert hierin Personen, Orte, Organisationen, Themen, Zahlen und Ereignisse. Die gesamte Datenbank ist kostenlos frei verfügbar, zur Analyse online oder zum Download.[218]
- Von Ontotext stammt das Angebot "**FactForge.net – Open Data and News about People, Organizations and Locations"**:[219] "*FactForge.net is a hub of Linked Open Data (LOD) and news articles about people, organizations and locations. It includes more than 1 billion facts from popular datasets such as DBpedia, Geonames, Wordnet, the Panama Papers, etc., as well as ontologies such as the Financial Industry Business Ontology (FIBO). It also includes a live stream of news articles and metadata linking news to entities and concepts: about 2000 articles/day tagged by Ontotext's Publishing platform*." Es ist also keine reine Nachrichtenseite, sondern verknüpft die Daten bereits mit vielfältigem Hintergrundwissen.[220] Die Speicherung als RDF-Graph ermöglicht eine Verknüpfung von Daten und somit die Generierung neuen Wissens. Abfragen sind beispielsweise über die Seite http://factforge.net/sparql möglich.

[216] https://www.gdeltproject.org/. GDELT bzw. GEG wird durch Google Jigsaw unterstützt.

[217] https://www.gdeltproject.org/.

[218] Vgl. https://www.gdeltproject.org/data.html. Die Daten nur des letzten Jahres umfassen aber bereits mehr als 2,5 TB.

[219] https://www.ontotext.com/knowledgehub/demoservices/factforge-explore-linked-open-data/.

[220] https://www.ontotext.com/knowledgehub/demoservices/factforge-explore-linked-open-data/.

3.2.8 Soziale Medien

Selbstverständlich können Alternative Daten auch aus den Sozialen Medien stammen.[221] Das bezieht sich einerseits auf Postings von Unternehmen, aber vor allem auch auf Posts durch die mit diesen verbundenen Personen. Es hängt von der jeweiligen Plattform ab, in welchem Umfang (Menge) bzw. mit welcher Datenrate (Geschwindigkeit) ein Abruf welcher Daten (Herkunft) möglich ist.

Einige Beispiele:

- LinkedIn - LinkedIn's Economic Graph[222]
- Facebook – Facebook Social Graph API[223]
- Instagram – Instagram Graph API[224]
- Twitter – Twitter APIs[225]

3.2.9 Lexika

Lexika bieten oft einen Hintergrund zur Einordnung und somit zum Verständnis von Daten. Neben das für menschliche Leser geschriebene Wikipedia treten auch Formen, die für eine automatisierte maschinelle Auswertung besser geeignet sind. Zu nennen sind insbesondere DBpedia[226] und Wikidata.

DBpedia enthält Informationen (u.a.) aus den Texten von Wikipedia und speichert sie in strukturierter Form ab. Ein Zugriff kann über eine Webanwendung, einen SPARQL-Endpunkt sowie durch einen Massendownload des Datenbestandes (Datenbankdump)[227] erfolgen. Die Inhalte ergeben einen Offenen Wissensgraphen (open knowledge graph; OKG).[228]

---

[221] Vgl. Russel / Klassen (2019).

[222] https://economicgraph.linkedin.com/.

[223] https://developers.facebook.com/docs/graph-api/.

[224] https://developers.facebook.com/docs/instagram-api.

[225] https://help.twitter.com/de/rules-and-policies/twitter-api.

[226] https://www.dbpedia.org/. Die Initiative ging von Deutschland aus.

[227] Unter https://dumps.wikimedia.org/.

[228] Vgl. https://www.dbpedia.org/about/.

Wikidata ist quasi eine Art Daten-Äquivalent zu den Textinhalten von Wikipedia. Ähnlich wie Wikipedia ist auch Wikidata frei bearbeitbar.[229] Mit Wikidata soll eine Wissensdatenbank geschaffen werden, deren Inhalte auch in anderen Wikimedia-Projekten, u.a. auch Wikipedia, genutzt werden können.[230] Alle Inhalte des Wissensgraphen sind von jedermann frei nutzbar.

3.2.10 Weitere Sammlungen

Im Internet finden sich weitere Sammlungen, die neben überwiegend nicht relevantem Material auch potentiell relevante Inhalte enthalten können.

Auf „**Linked Open Data Cloud**"[231] wird auf Datenbestände verwiesen, die gemäß der Linked Data Prinzipien publiziert wurden, also als RDF-Graph.[232] Der Datensatz muss mindestens 1000 Einträge umfassen, vollständig abrufbar sein (durch RDF Crawling, einen Massendownload oder über einen SPARQL Endpunkt) sowie mit dem bisherigen Datenbestand verlinkt werden.

Für alternative Daten aus der klassischen Sicht von Finanzinvestoren gibt es eine Vielzahl **kommerzieller Datenlieferanten**. Einen Einstieg geben beispielsweise ravenpack.com,[233] alternativedata.org[234] oder datarade.ai.[235]

Eine Vielzahl von anderen bereits im Web gecrawlte Daten wird frei über „commoncrawl.org" zur Verfügung gestellt.[236]

---

[229] Vgl. https://www.wikidata.org/wiki/Wikidata:Introduction.

[230] https://www.wikidata.org/wiki/Wikidata:Main_Page.

[231] https://lod-cloud.net/.

[232] Vgl. https://www.w3.org/DesignIssues/LinkedData.html.

[233] https://www.ravenpack.com.

[234] https://alternativedata.org/data-providers/.

[235] https://datarade.ai/data-categories.

[236] http://commoncrawl.org/.

Zuletzt genannt sei **Dataset Search**, die Datensuchmaschine von Google.[237] Dabei handelt es sich – im Unterschied zur „normalen“ Google Suche - um eine auf Datensammlungen spezialisierte Suchmaschine.[238] Ihre Nutzer können durch einfache Suchbegriffe Datensätze finden, die in Repositories im weltweiten Web gehostet werden.[239]

3.3 Ergebnis und Ausblick

Die Auflistung Alternativer Datenquellen belegt, dass bereits heute ein umfangreiches Angebot an automatisiert auswertbaren Datenbeständen vorliegt, die sich zur Verprobung von und Verknüpfung mit steuerlichen Daten eignen.

Dieses Flickwerk an Daten ist beileibe nicht vollständig und nicht für alle Bereiche geeignet. Jedoch lassen sich einige Trends feststellen: Der Umfang der Daten wächst beständig und zwar sowohl in der Tiefe (durch vermehrte Einträge in bestehende Datensätze) als auch in der Breite (durch zusätzliche Datensätze).

Ferner werden die bisherigen „Dateninseln“ immer stärker miteinander verknüpft, was deren Auswertung enorm vereinfacht. So hat die EU als Ziel den Aufbau eines gemeinsamen europäischen Datenraums ausgegeben.[240] Das geschieht einerseits innerhalb einzelner Plattformen, andererseits durch wechselseitige Verweise insbesondere mit Techniken aus dem Internet.

Zusätzlich gibt es immer mehr Offene Daten in Deutschland,[241] in der Europäischen Union und weltweit. Das data.europa.eu-Portal bietet einen Zugang zu offenen Daten aus internationalen, EU-, nationalen, regionalen und lokalen Datenportalen.[242]

---

[237] https://datasetsearch.research.google.com/.

[238] Ähnlich wie Google Scholar wissenschaftliche Veröffentlichungen sucht. Siehe https://scholar.google.com/.

[239] Vgl. https://datasetsearch.research.google.com/help.

[240] EU, Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen „Aufbau eines gemeinsamen europäischen Datenraums“ vom 25.4.2018, COM(2018) 232 final.

[241] Zu Offenen Daten im Inland am Beispiel von Berlin: Berliner Senat, Verordnung zur Bereitstellung von allgemein zugänglichen Datenbeständen (Open Data) durch die Behörden der Berliner Verwaltung (Open Data Verordnung - OpenDataV) vom 24. Juli 2020; Berlinonline (2021): Das Berliner Open-Data-Handbuch, https://berlinonline.github.io/open-data-handbuch/ (13.3.2021).

[242] Vgl. https://data.europa.eu/en/about/about-dataeuropaeu. Es ersetzt das EU Open Data Portal und das European Data Portal.

Daher werden sich die Möglichkeiten zur Nutzung (auch) Alternativer Daten künftig weiter verbessern.

4 Prototypische Umsetzung

4.1 Technische Basis

4.1.1 Aufgabenstellung

Für den Einsatz in der Finanzverwaltung besteht die Aufgabe nicht nur in der Sammlung von Alternativen Daten. Diese müssen vielmehr in einer Form gespeichert werden, die gezielte Abfragen auf den integrierten Gesamtdatenbestand ermöglicht. Daraus ergeben sich folgende Arbeitsschritte:

- Sammlung der Alternativen Daten als Rohdaten
- Überführung in ein einheitliches Datenformat
- Integration der Teil-Datenbestände
- Erforschung des Gesamtdatenbestands durch Abfragen

4.1.2 Datensammlung

Die Alternativen Daten können grundsätzlich in ganz verschiedenen **Dateitypen** vorliegen. Hierzu zählen u.a. einfache TXT-Dateien, PDF, MS-Word, MS-Excel, CSV, Bilder/Scans als JPEG oder TIFF,[243] HTML, XML, XBRL, iXBRL, JSON oder RDF-Serialisierungen.

Diese Vielfalt lässt sich technisch gesehen geordnet in einem sog. Data Lake speichern, der sowohl strukturierte wie auch semi-strukturierte und unstrukturierte Daten in ihrem Raw-Format aufnehmen kann. Dieser große See an Rohdaten „verschluckt" somit zunächst einmal alles. Die künftige Verwendung kann dabei vorläufig offenbleiben.

In vielen Fällen wird den deutschen Behörden **kein kompletter Datensatz** zur Verfügung stehen. Der Zugriff erfolgt durch Abfragen, die sich nach konkreten Suchobjekten (Unternehmen, Personen) richten. Idealerweise ist eine **Programmschnittstelle – API** (Application Program Interface) – vorhanden.

[243] Bilder, aber auch Videos oder Audios werden nicht als alphanumerische Zeichen gespeichert, sondern in Binärformaten. Man spricht auch von Blobs (Binary Large Objects).

Ist eine solche nicht verfügbar, so sind Abfragen (nur) über die Weboberfläche möglich. Solche Abfragen müssen im Einzelfall erstellt und ausgewertet werden, was mit viel Arbeit verbunden ist.

Jedoch gibt es hier ein sehr hohes Automatisierungspotenzial durch sog. „**Robotic Process Automation (RPA)“**. Unter Robotic Process Automation darf man sich keinen körperlichen Roboter vorstellen. Es ist vielmehr ein Programm, das Anstelle des Menschen den Computer und die dort installierte Software über die Benutzeroberfläche bedient. *Moffitt / Rozario / Vasarhelyi* beschreiben diese Vorgehensweise sehr anschaulich wie folgt: „First and foremost, RPA robots conduct work the same way that humans do, through the software presentation layer. Logins, emails, analyses, report building, data entry, and other functions are still completed. RPA robots can be compared to the recorded macros in Excel that automate specific tasks. The primary difference between the two is that RPA ‘macros’ can be recorded to work with virtually any existing desktop or server software. RPA software generally includes an interface with a record button that, when activated, generates a script, or robot, as a user performs the task that is to be automated. With some configuration, robots can be trained to read emails, open PDFs, identify salient information, enter data into ERP systems, and send an email to specific supervisors when ambiguity or errors are encountered.”[244] Der von der RBA-Software erstellte Programmcode wird auch als „Robot“ oder kurz „Bot“ bezeichnet.

RPA eignet sich besonders für wohldefinierte Aufgaben.[245] Hierzu zählen etwa Abfragen ausländischer Register oder Abfragen kommerzieller Datensammlungen (LinkedIn, PEP etc.). In solchen Anwendungsszenarien müssten folgende Bedingungen für den Einsatz von RPA vorliegen:

- Die einzelne Abfrage ist kostenlos oder die mit dem Abruf verbundenen Kosten werden in Kauf genommen. Das könnte beispielsweise bei einer Flatrate von monatlich 1.000€ der Fall sein, wohingegen Kosten von 5€ pro Zugriff schwer kalkulierbar sind.

[244] Moffitt / Rozario / Vasarhelyi (2018), S. 2.

[245] Vgl. Huang / Vasarhelyi (2019).

- Es gibt keine Mechanismen, die eine Abfrage durch Bots (anstelle von Menschen) zu verhindern suchen oder diese Mechanismen werden bewusst überwunden. Hierzu werden sog. CAPTCHA-Tests eingesetzt. CAPTCHA steht für „Completely Automated Public Turing test to tell Computers and Humans Apart“.[246] Das kann beispielsweise in Form von 9 oder 12 Fotos geschehen, auf denen der Nutzer diejenigen markieren soll, auf denen Fahrräder (oder Ampeln, Zebrastreifen etc.) zu sehen sind. Auch das deutsche Unternehmensregister versucht Bots auszuschließen.

4.1.3 Einheitliches Datenformat

Nach erfolgter Sammlung von Rohdaten müssen die heterogenen Datenbestände in der Folge transformiert werden, um sie besser analysieren zu können. Für die Transformation in ein einheitliches Datenformat sollte dieses gewissen Bedingungen genügen.

In jedem Fall muss es sich um ein **strukturiertes Format** handeln. Gegebenenfalls müssen vorher unstrukturierte Daten in strukturierte Daten überführt werden:

- Im Extremfall liegen Scans von Dokumenten in Binärform wie z.B. JPEG vor. Über diese Bilddaten muss dann eine sog. OCR-Software (Optical Character Recognition) laufen, um die abgebildeten Buchstaben und Zahlen zu identifizieren und alphanumerisch in einer TXT-Datei abzuspeichern.
- Aber auch Fließtexte müssen noch inhaltlich analysiert werden. Ein Text enthält beispielsweise Aussagen zu Unternehmen („ABC GmbH“), Personen („Max Müller“), Funktionen („Prokurist“), Orten („Luxembourg“) usw. Diese sollen erkannt und extrahiert werden. Hierbei hilft eine sog. „Named Entity Recognition (NER)“-Software.[247]

Die Struktur der Daten sollte **wesentliche Eigenschaften der interessierenden Objekte** (Entitäten, Entities) erfassen. Zu den Objekten gehören insbesondere Unternehmen und Personen. Sie werden durch verschiedene Eigenschaften (Attribute) sowie ihre Beziehungen (Relationships) untereinander charakterisiert.[248]

---

[246] Vgl. Cloudflare (2021).

[247] Vgl. Balogh (2018). Andere Aufgabe wären z.B. das Erkennen von Personen auf Bildern oder in Videos.

[248] Vgl. Balogh (2018), S. 3f.

Das Datenmodell sollte **einfache Erweiterungen** des ursprünglichen Datenbestandes erlauben. Darunter ist weniger das Hinzufügen einer zusätzlichen Anzahl derselben Daten gemeint, also wenn sich etwa die Liste von „Politisch Exponierten Personen" um weitere Personen erweitert. Es geht vielmehr um das Auftreten von Daten mit neuen, zusätzlichen Merkmalen. Beispielsweise können die Aktivitäten eines Tochterunternehmens auch um das Merkmal ergänzt werden, ob es sich um Einkünfte aus passiven Betätigungen im Sinne von § 8 AStG handelt. Daten haben häufig Lücken, weil nicht in allen Fällen sämtliche Eigenschaften bekannt sind.

Das Datenformat soll schließlich eine **problemlose Verknüpfung** der unterschiedlichen Datenquellen erlauben. Dabei ist die Verwendung etablierter und erprobter **Standards** förderlich.

Vor dem Hintergrund dieser Anforderungen gibt es eine Reihe von Datenbank-Konzepten, die geeignet sind, strukturierte Daten zu speichern.[249]

(1) Heute am weitesten verbreitet sind **relationale Datenbankmanagementsysteme (RDBMS)**,[250] die Daten aus logischer Sicht in Tabellen organisieren. Die Tabellen bestehen aus vertikalen Spalten (Datenfelder mit festgelegter Datenstruktur) und horizontalen Zeilen (Tupel oder Datensätze). Der gesamte Tabellenentwurf wird als Datenbankschema bezeichnet. Dieses vordefinierte Schema bedeutet ein starres Datenmodell, welches im Vorfeld ein sorgfältiges Design erfordert, denn jede zukünftige Änderung des Datenschemas ist aufwändig und kostspielig. Die Definition der Datenstruktur, Bearbeitungen (Einfügen, Verändern, Löschen) und Abfragen erfolgen über die Sprache SQL.

(2) „**Document Stores**" sind hingegen Datenmodelle,[251] die auch für die Verwaltung von unstrukturierten (z. B. Text) oder halbstrukturierten (z. B. XML) Daten geeignet sind und in der Regel eine hierarchische Natur aufweisen.[252] Bei dokumentenorientierten Datenbanken können die in jedem Dokument gespeicherten Daten unterschiedliche

[249] Zu anderen Formen von Datenbankmanagementsystemen vgl. die Technik- und Marktübersicht auf https://db-engines.com.

[250] Zu den relationalen Datenbankmanagementsystemen gehört beispielsweise MySQL. Vgl. https://www.mysql.com., https://db-engines.com/de/ranking/relational+dbms.

[251] Das beliebteste dokumentenorientierte Datenbankmanagementsystem ist etwa MongoDB. Vgl. https://www.mongodb.com, https://db-engines.com/de/ranking/document+store.

[252] Vgl. Foote (2018); Nayak / Poriya / Poojary (2013); Sharma / Dave (2012).

Datenfelder und Datenfeldtypen haben. Dies ist bei Datensätzen in relationalen Datenbanken nicht möglich. Das macht dokumentenorientierte Datenbanken flexibler bei Änderungen. Umgekehrt ist deshalb eine Analyse des Gesamtdatenbestands durch Abfragen nicht auf einfache Weise möglich.

(3) Einen Ausweg können **Graphen-Datenbanken** bieten. Graphendatenmodelle basieren auf der mathematischen Graphentheorie. Ganz allgemein speichern sie Informationen über Objekte (Entities) und deren Beziehungen (Relations). Die Objekte werden als Knoten und ihre Beziehungen als Kanten gespeichert. Daher zeichnen sich Graphen-Datenbanken durch die Speicherung und Verarbeitung stark miteinander verbundener Datenstrukturen aus, im Unterschied zum relationalen Datenmodell, bei dem die Tabellen nur lose verbunden sind. Dies passt sehr gut zur Aufgabe einer Verifizierung von steuerlichen Daten durch (ggf. eine Kette von) Alternativen Daten. Neue Knoten und Kanten können zu einer bestehenden Graphen-Datenbank hinzugefügt werden, sobald man sie benötigt, wodurch die Datenbank auch langsam organisch wachsen kann.

Graphen-Datenbanken haben in jüngster Zeit an Beliebtheit gewonnen, eben weil sie heterogene Daten, Integration neuer Datenquellen und ein für die Analytik geeignetes Strukturschema ermöglichen.[253] Beispiele sind die Nutzung von Graphen-Datenbanken u.a. durch Google,[254] Facebook, LinkedIn, DBPedia, Wikidata, FactForge, OCCRP Aleph oder das ICJI.[255]

Im Wesentlichen gibt es zwei Hauptmodelle von Graphen-Datenbanken: **Labeled-Property Graph** und **RDF Stores**.[256] Der prominenteste Vertreter des Labeled-Property Graph ist Neo4j,[257] zu den RDF Stores gehören u.a. GraphDB von Ontotext (kommerziell) und „Apache Jena – TDB“ (Open Source).[258]

Wegen der Tatsache, dass es keinen Standard für das Labeled-Property Graph-Datenmodell gibt, existiert leider auch keine gemeinsame Abfragesprache für den Zugriff

[253] Vgl. Polikoff (2020).

[254] Vgl. Singhal (2012) und https://support.google.com/knowledgepanel/answer/9787176?hl=de.

[255] Siehe die jeweiligen Hinweise auf die Wissensgraphen weiter oben im Text.

[256] Vgl. Barrasa (2017), Polikoff (2020).

[257] Vgl. https://db-engines.com/de/article/Graph+DBMS.

[258] Vgl. https://db-engines.com/de/article/RDF+Stores.

auf Daten in der Graphen-Datenbank sowie kein Standard zum Importieren von Daten (Upload) in die Graphen-Datenbank. Daher ist es Sache des DBMS-Anbieters, seine eigene Abfragesprache oder Vorgaben für den Datenaustausch zu definieren. Beispiele sind Cypher,[259] Gremlin, GraphQL und andere proprietäre Sprachen.[260]

Für den angestrebten Einsatzzweck haben Graphen-Datenbanken als RDF Store entscheidende Vorteile, denn sie beruhen auf anerkannten Standards.

So gehört RDF zu den **Standards der EU** und bildet einen Teil der Digitalisierungsstrategie. RDF wurde am 20. Juli 2017 von der EU als offizieller technischer Standard (RTS) für Informations- und Kommunikationstechnologien (IKT) angenommen (ähnlich wie XBRL als RTS für IKT am 28. Januar 2016).

RDF ist Teil der internationalen **Standards des W3C** (World Wide Web Consortium) für das Internet:

> *"RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.*
>
> *RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.*
>
> *This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This graph view is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations."*[261]

[259] Cypher wurde inzwischen von Neo4j als Open Source frei verfügbar gemacht. Des Weiteren existieren Bemühungen, einen einheitlichen GQL Standard (Graph Query Language Standard) zu etablieren. Vgl. https://www.gqlstandards.org/.

[260] Vgl. https://www.gqlstandards.org/existing-languages.

[261] https://www.w3.org/RDF.

Zu den weiteren relevanten Standards des W3C gehören (u.a.) das RDF Schema (RDFS),[262] die Web Ontology Language (OWL)[263] sowie die Abfragesprache SPARQL.[264]

Als **Zwischenergebnis** lässt sich festhalten, dass Alternative Daten am sinnvollsten in einem **Wissensgraphen (Knowledge Graph) als RDF Store** nach den Standards des W3C gespeichert werden sollten, da dies den Anforderungen am besten entspricht. Hierfür müssen auch keine Lizenzierungskosten anfallen, da alle benötigten Bestandteile als Offene Standards und Freie Software verfügbar sind.

Ein Wissensgraph grenzt sich folgendermaßen von anderen Datensammlungen ab: „*A knowledge graph (1.) mainly describes real world entities and their interrelations, organized in a graph, (2.) defines possible classes and relations of entities in a schema, (3.) allows for potentially interrelating arbitrary entities with each other and (4.) covers various topical domains."*[265]

Wie von Kejriwal beschrieben, kann ein Knowledge Graph als eine „graphentheoretische Darstellung menschlichen Wissens beschrieben werden, so dass es von einer Maschine mit **Semantik** aufgenommen werden kann".[266] Das Journal of Web Semantics gibt folgende Definition an: "Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities."[267]

Unter Semantik versteht man grundsätzlich die Lehre von der Bedeutung von Dingen. Menschen haben bereits eine gewisse Vorstellung, was bestimmte Dinge (z.B. „Geschäftsführer", „Anteilseigner", „Sitz", „Tochtergesellschaft") bedeuten. Eine Software besitzt das erst einmal nicht. Daher müssen ihr die nötigen Bedeutungen explizit und auf formale Weise mitgeteilt werden. Hierin besteht die Relevanz von sog. Ontologien, welche in Wissensgraphen verwendet werden.

---

[262] https://www.w3.org/2001/sw/wiki/RDFS.

[263] https://www.w3.org/2001/sw/wiki/OWL. Näheres in Uschold, (2018).

[264] https://www.w3.org/2001/sw/wiki/SPARQL. Näheres siehe weiter unten unter 5.

[265] Vgl. Paulheim (2017).

[266] Vgl. Kejriwal (2019).

[267] Vgl. Kroetsch / Weikum (2016). https://www.websemanticsjournal.org/index.php/ps/ announcement/view/19.

Ontologien sind formale semantische Datenmodelle, die für unsere Domäne (d.h. den uns interessierenden Bereich) die Arten von existierenden Dingen (Objekten, Entitäten) und die Eigenschaften, die zu deren Beschreibung verwendet werden können, definieren.

Um die im ersten Schritt aus verschiedensten Quellen zusammengetragen Daten zu vereinheitlichen, sind Ontologien als essenzieller Bestandteil des Prozesses anzuwenden.[268]

Ontologien stellen in der Informatik das Grundgerüst einer fehlerlosen und eindeutigen Kommunikation dar. Sie sind unabhängig von einer bestimmten Anwendungssoftware und lassen sich daher mehrfach einsetzen.[269] Formale Ontologien sollen semantische Strukturen für alle Teilnehmer verständlich machen.[270] In diesem Zusammenhang sind Ontologien daher besonders relevant, um Wissen weiterzugeben und zu verarbeiten.[271]

Ein Wissensgraph integriert Daten aus verschiedenen Quellen in einer Graphenstruktur mit Hilfe von Ontologien, um anhand dieser Strukturen Wissen zu modellieren.[272] Wissensgraphen werden durch sog. **Triple** ausgedrückt. Ein Triple erlaubt es, Beziehungen (= Kante) zwischen zwei Entitäten (=Knoten) durch drei Elemente auszudrücken:

(1) Subjekt-Entität, (2) Prädikat und (3) Objekt-Entität.

Abb. 13: Beispiel für ein RDF-Triple

Ein komplexeres Beispiel für einen Wissensgraphen mit Tripeln aus jeweils zwei Knoten und einer Kante findet sich in der Abbildung 4.

[268] Vgl. Fensel et al. (2020); Kendall / McGuinness (2019).

[269] https://it-talents.de/it-wissen/ontologie-in-der-informatik/.

[270] Vgl. Allemang / Hendler (2011), S. 13.

[271] Vgl. Obitko 2007.

[272] https://www.turing.ac.uk/research/interest-groups/knowledge-graphs.

4.1.4 Datenintegration

Die Teil-Datenbestände, welche in einen RDF-Graphen überführt wurden, leiden jedoch in der Regel darunter, dass zur Identifikation der individuellen Unternehmen und Personen in den Teil-Datenbeständen unterschiedliche Schlüssel (Identifiers) verwendet wurden.

Wie weiter oben beschrieben kommen u.a. in Frage:

- Handelsregister-Nummer in Verbindung mit dem Registergericht
- entsprechende Nummern aus den Registern andere Länder
- (deutsche) Steueridentifikationsnummer
- LEI
- PermID
- QID Wikidata-Kennung[273]
- Open Corporates Company Number
- usw.

Zur Schaffung eines Gesamt-Datenbestandes aus den heterogenen Datenbeständen müssen die unterschiedlichen Identifiers (Schlüssel) der Objekte in einzelnen Datenbeständen mit sog. Mappings untereinander verknüpft werden. Darunter versteht man etwa eine Zuordnung, dass

- das Unternehmen mit der LEI 1234568ABCD12345678 und
- das Unternehmen mit der Handelsregister-Nummer Fürth_HRB12345

identisch sind.

Hierzu kann in Wissensgraphen die OWL-Eigenschaft „SameAs“ verwendet werden, wie das folgende Beispiel für Personen (anstelle von Unternehmen) zeigt:[274]

“*The built-in OWL property* `owl:sameAs` *links an individual to an individual. Such an owl:sameAs statement indicates that two URI references actually refer to the same thing: the individuals have the same ‘identity’.*

[273] Vgl. https://www.wikidata.org/wiki/Q43649390.

[274] https://www.w3.org/TR/owl-ref/#sameAs-def, Abschnitt 5.2.1.

For individuals such as 'people' this notion is relatively easy to understand. For example, we could state that the following two URI references actually refer to the same person:

```
<rdf:Description rdf:about="#William_Jefferson_Clinton">
  <owl:sameAs rdf:resource="#BillClinton"/>
</rdf:Description>
```
"

Aus logischer Sicht werden dann die beiden Einträge „William_Jefferson_Clinton" und „BillClinton" als ein und dieselbe Person behandelt, auch wenn sie in getrennten Datensätzen gespeichert wurden. Das ist selbstverständlich auch für verschieden erfasste Unternehmen möglich, also z.B. „BMW", „Bayerische Motoren Werke Aktiengesellschaft" oder „Bayerische Motoren Werke AG".

Eine Identitätsbeziehung kann auch bei gleichzeitiger Nutzung mehrerer Klassifikationsschemata (Ontologien) definiert werden. Beispielweise wird die Sportart „Fußball" in England „Football", in den USA „Soccer" genannt. Für die Fußballmannschaft könnte man definieren: [275]

"*In OWL Full ... we can use the* `owl:sameAs` *construct to define class equality, thus indicating that two concepts have the same intensional meaning. An example:*

```
<owl:Class rdf:ID="FootballTeam">
  <owl:sameAs rdf:resource="http://sports.org/US#SoccerTeam"/>
</owl:Class>
```
"

Eine Verknüpfung verschiedener Datenbestände wird auch in der EU angestrebt. Beispielsweise handelt es sich beim **euBusinessGraph**[276] um ein Forschungsprojekt von Horizon 2020.[277] Ziele waren:[278]

Ein System von Identifikatoren (Schlüsseln, Identifiers) für unternehmensbezogene Daten und Entitäten.

[275] https://www.w3.org/TR/owl-ref/#sameAs-def, Abschnitt 5.2.1.

[276] https://www.eubusinessgraph.eu/.

[277] Fördernummer Grant Agreement No. 732003.

[278] Vgl. https://www.ontotext.com/knowledgehub/current/eubusinessgraph/.

Dies soll die Verknüpfung von Daten zwischen Entitäten in verschiedenen Ländern und über unterschiedliche Sprachen hinweg ermöglichen. Es bildet eine Grundlage für die Verknüpfung von Daten. Dazu wurden in Europa verwendete Identifikatoren genutzt und Mappings erstellt.

Entwicklung gemeinsamer mehrsprachiger Datenmodelle (Ontologien und Vokabulare) für Unternehmensdaten. Als Grundlage dienten vorhandene Schemata/Vokabulare/Ontologien wie EU Core Vocabs[279] (W3C Org, W3C RegOrg, W3C Location, Person - nicht W3C), schema.org,[280] GLEI, FIBO,[281] Wikidata,[282] ADMS.[283, 284] Das Datenmodell ist Open Source und auf GitHub verfügbar.[285]

4.1.5 Erforschung des Gesamtdatenbestands

Nach der Implementierung des generierten Graphen in eine Datenbank kann der bestehende Graph abgefragt werden, um Informationen zu extrahieren. SPARQL (SPARQL Protocol and RDF Query Language) stellt die Abfragesprache für RDF-Daten im Semantic Web dar. Mittels dieser können komplexe Abfragen von RDF-Datensätze im Linked Open Data Format durchgeführt werden, um die spezifischen Informationen aus diesen herauszuziehen.[286]

Wie schon oben erwähnt, ist einer der Hauptvorteile des Graphen-Datenmodells die Fähigkeit, in Abfragen Daten zu verbinden, Korrelationen zwischen verbundenen Daten zu erkennen und neues Wissen aus den bisher gespeicherten Daten abzuleiten (sog. Inferenz). Diese Inferenz bzw. semantische Generierung von neuem Wissen fügt daher der Verwendung eines einfachen Graphen eine neue „Wissens"-Ebene hinzu, die ihn in einen Wissensgraphen verwandelt.

---

279 https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/e-government-core-vocabularies/release/20#download-links.

280 https://schema.org/.

281 Financial Industry Business Ontology (FIBO), vgl. https://spec.edmcouncil.org/fibo/ontology/master/latest/.

282 https://www.wikidata.org.

283 Asset Description Metadata Schema (ADMS), vgl. https://www.w3.org/TR/vocab-adms/.

284 Vgl. https://www.eubusinessgraph.eu/eubusinessgraph-ontology-for-company-data/,

285 https://github.com/euBusinessGraph/eubg-data. Zur Erläuterung Alexiev (2019).

286 Vgl. Auer / Pellegrini / Sack (2014).

Ein Beispiel dafür zeigt Abbildung 14. Es ist die Graphen-Darstellung eines örtlichen Coffee House („My_Local_Cafe“, My_Local_Cafe), der Eigentümer-Struktur und der Standorte.

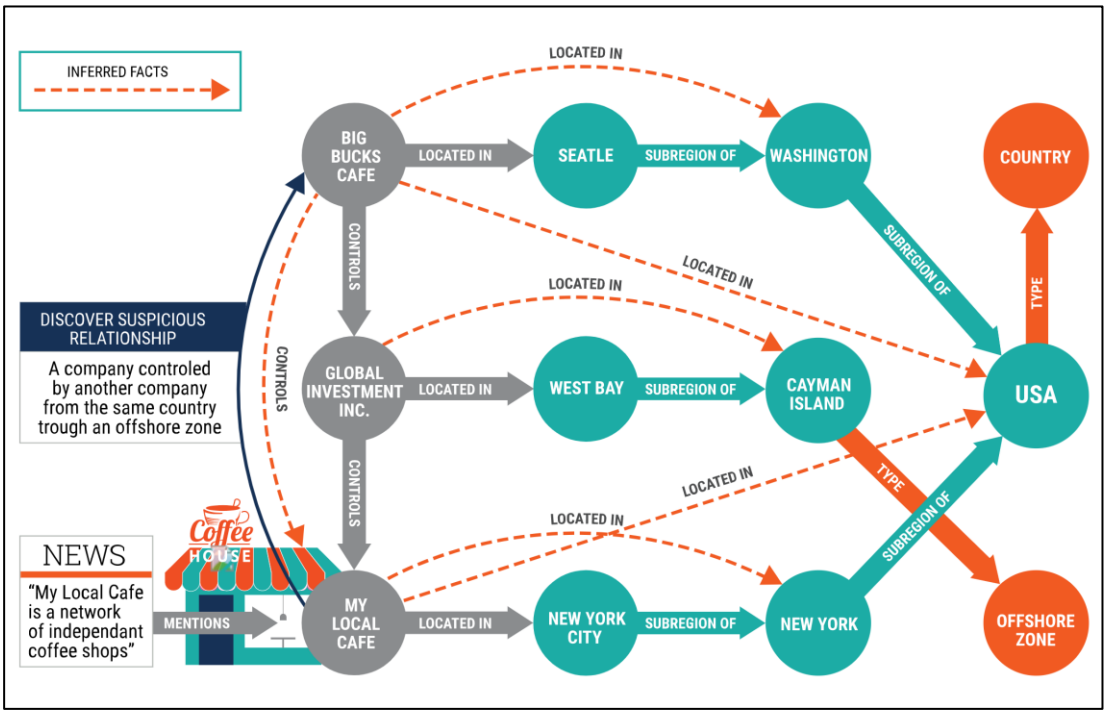
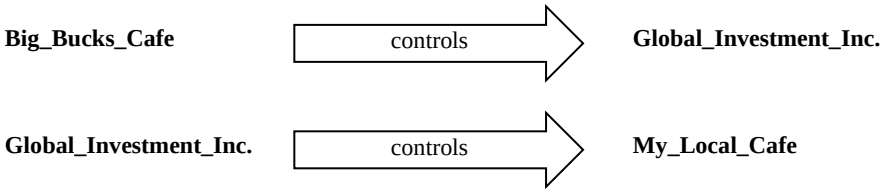


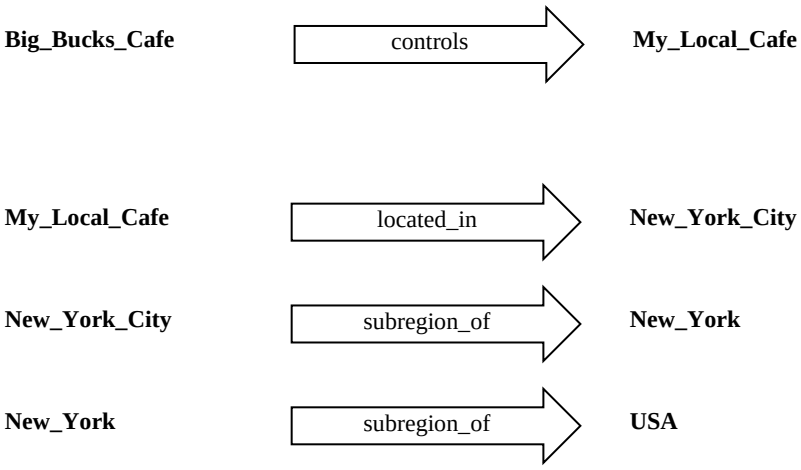
Abb. 14: Erkennen steuerlich risikobehafteter Unternehmensstrukturen
Quelle: Ontotext, https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/ (25.9.2021)

Wenn man das Netz mit seinen verschiedenen Knoten und deren Beziehungen (Kanten) untereinander durchquert, so könnte man auf eine steuerlich verdächtige Beziehung der folgenden Art schließen: „*Ein Unternehmen, das von einem anderen Unternehmen aus demselben Land indirekt über eine Offshore-Zone kontrolliert wird*“.

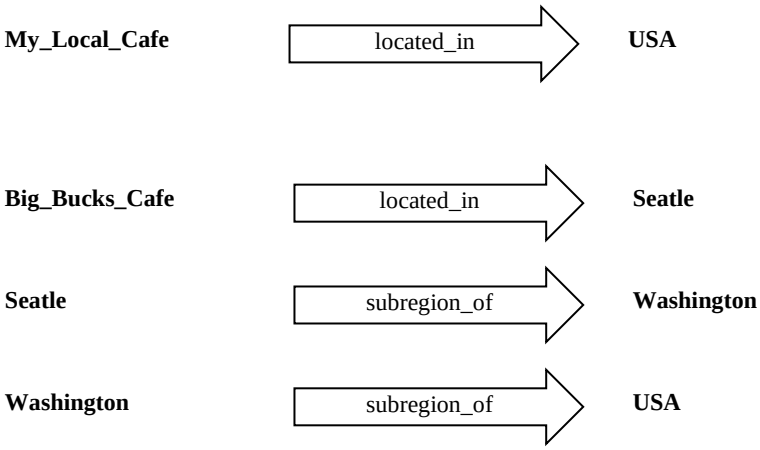
Diese Inferenzkette baut sich folgendermaßen auf:



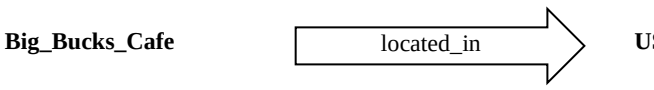
Erste Inferenz:

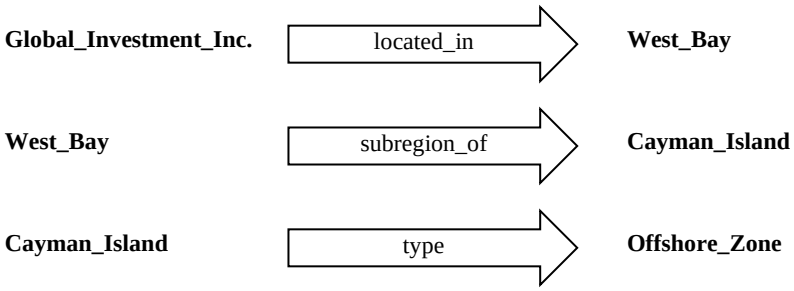


Zweite Inferenz:

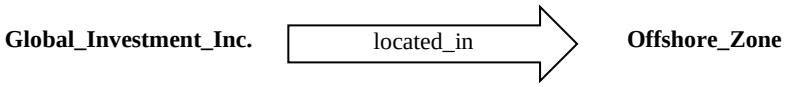


Dritte Inferenz:

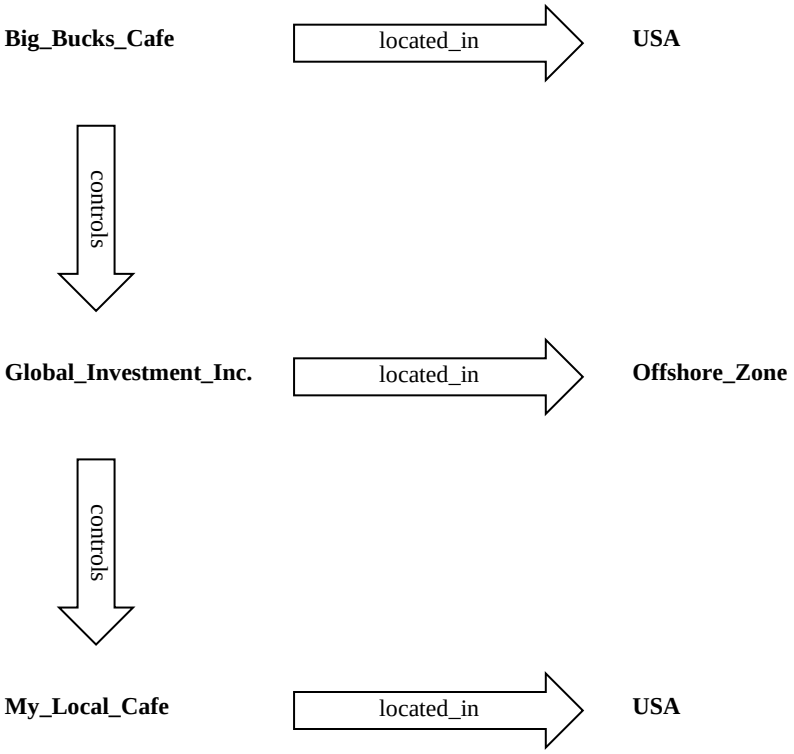




Vierte Inferenz:



Zusammen erhält man das folgende „verdächtige“ Muster:



Mithilfe zusätzlicher Konvertierung, je nach verwendetem Tool (z.B. Gephi), können auch komplexere Graphen automatisch visualisiert werden, um weitere Analysen zu ermöglichen. In diesen wird aus dem Subjekt und dem Objekt eines Triples ein Knoten und aus dem Prädikat eine Kante.[287] Aufgrund der Visualisierung können dadurch auffällige Muster, z.B. im Zuge der Betrugsaufdeckung, erkannt und dargestellt werden.

[287] Arnaout / Elbassuoni (2018).

4.2 Codierungsbeispiele

4.2.1 Deutsches Unternehmensregister

Wie beschrieben verfolgt das Projekt „offeneregister.de“ das Ziel, die faktischen Zugriffsbeschränkungen des offiziellen Handelsregisters zu beseitigen, indem der Inhalt des Handelsregisters aus öffentlich zugänglichen Quellen, insbesondere aus gesetzlich vorgeschriebenen Handelsregisterbekanntmachungen,[288] rekonstruiert wird.

Der Gesamtdatenbestand ist hier frei verfügbar. Der Download kann als JSON-Datei oder als SQLite-Datenbank-Datei erfolgen. Es wurde die zweite Form gewählt. Die gezippte Datei hatte ein Volumen von ca. 755 MB. Entpackt umfassen die Daten ca. 2,6 GB. Der Download erfolgt im Februar 2021.[289] Damit ergeben sich eine Reihe interessanter Auswertungsmöglichkeiten, die mit dem offiziellen Register so nicht möglich wären.

Von den enthaltenen Daten wurden zwei Tabellen genutzt, die mit „company“ und „officer“ bezeichnet werden. Company sind die im Handelsregister eingetragenen Unternehmen. Entgegen dem Anschein dieser Bezeichnung handelt es sich selbstverständlich nicht nur um Kapitalgesellschaften. Officer umfasst die vertretungsberechtigten Personen, die entweder natürliche oder jursitische Personen sein können.

Der Ausgangsdatenbestand von company umfasst 5.305.727 Zeilen und von officer 4.803.514 Zeilen. Jedoch sind hierin sowohl nicht mehr aktive Unternehmen wie auch aktuell nicht mehr vertretungsberechtigten Personen enthalten.

Dies auszuschließen und sich zugleich auf die wesentlichsten Merkmale zu beschränken, wurden durch SQL-Befehle zwei sog. Views (v_company, v_officer) auf die beiden Ausgangstabellen definiert:

[288] Vgl. handelsregisterbekanntmachungen.de.

[289] Wobei der Datenbestand offensichtlich nicht permanent aktualisiert wird, denn laut Beschreibung handelt es sich um eine Version der Daten vom 15.2.2019. Das ändert aber nichts an den dargestellten grundsätzlichen Auswertungsmöglichkeiten.

```
CREATE VIEW IF NOT EXISTS v_company
AS
SELECT
id as c_id,
company_number as c_number,
name AS c_name,
registered_office||'_'||registered_address as c_address,
native_company_number
FROM company
WHERE
current_status = "currently registered";
```

Damit verblieben 3.341.017 aktive Unternehmen. Einige Datensätze lauten beispielsweise:

c_id	c_number	c_name	c_address	native_company_number
1	K1101R_HRB150148	olly UG (haftungsbeschränkt)	Hamburg_Waidmannstraße 1, 22769 Hamburg.	Hamburg HRB 150148
2	R1101_HRB81092	BLUECHILLED Verwaltungs GmbH	Düsseldorf_Oststr.	Düsseldorf HRB 81092
4	R1101_HRB45109	Albert Barufe GmbH	Hilden_Hans-Sachs-Straße 11, 40721 Hilden.	Düsseldorf HRB 45109
7	K1101R_HRB82839	Verwaltung IFÖ Zweite Immobilienfonds für Österreich GmbH	Hamburg_Königstraße 28, 22767 Hamburg.	Hamburg HRB 82839
5	R1101_HRB37996	ITERGO Informationstechnologie GmbH	Düsseldorf_ERGO-Platz 1, 40477 Düsseldorf.	Düsseldorf HRB 37996

Für den View auf officer gilt:

```
CREATE VIEW IF NOT EXISTS v_officer
AS
SELECT
id as o_id,
name AS o_name,
position as o_position,
type as o_type,
company_id as c_number,
city as o_city
FROM officer
WHERE
dismissed ISNULL
```

Es verblieben 3.341.017 „officers“. Ein Auszug aus v_officer:

o_id	o_name	o_position	o_type	c_number	o_city
1	Oliver Keunecke	Geschäftsführer	person	K1101R_HRB150148	Hamburg
2	Christof Wessels	Geschäftsführer	person	R1101_HRB81092	Cloppenburg
3	Christof Wessels	Geschäftsführer	person	R1101_HRB81092	Cloppenburg
11	Joachim Wehrkamp	Geschäftsführer	person	H1101_H1101_HRB18423	Thedinghausen
12	Jörn-Michael Gauss	Geschäftsführer	person	H1101_H1101_HRB18423	Bremen

Auf einen Import in einen RDF Store konnte verzichtet werden, da die wechselseitige Verknüpfung vom Typ „Person – istOfficerBei – Unternehmen“ bzw. „Unternehmen – hatOfficer – Person“ sehr begrenzt sind und sich noch einfach in SQL abbilden lassen.

Im obigen Auszug aus v_officer sind nur natürliche Personen als Geschäftsführer enthalten. Einen Überblick zu den Möglichkeiten und ihrer Häufigkeit (Anzahl der Fälle) gewinnt man durch folgende Datenbankabfrage:

```
SELECT o_position, o_type, count(o_position), count(o_type)
FROM v_officer
GROUP BY o_position, o_type
```

Die Ergebnistabelle zeigt:

o_position	o_type	Anzahl
Geschäftsführer	company	61
Geschäftsführer	person	2192235
Inhaber	company	238
Inhaber	person	126764
Liquidator	company	11990
Liquidator	person	313533
Persönlich haftender Gesellschafter	company	276621
Persönlich haftender Gesellschafter	person	88173
Prokurist	company	4
Prokurist	person	328938
Vorstand	person	2460
Summe		**3341017**

Interessante Einblicke lassen sich durch komplexere Abfragen gewinnen.

Unter einer sog. „Briefkastenfirma“ versteht man meist ein Unternehmen, das an seiner Adresse keine eigenen Büroräume hat und keine eigenen Mitarbeiter beschäftigt. Die physische Präsenz beschränkt sich somit auf einen Briefkasten bzw. ein Postfach.

Grundsätzlich sind solche Briefkastenfirmen nichts Illegales. Grund hierfür kann die gesellschaftsrechtliche Verkörperung von einzelnen Vermögenswerten wie Grundstücken, Schiffen, aber auch Patenten und anderen Rechten sein. Eine solche Gesellschaft – in der reduziertesten Form eben als Briefkastenfirma – ermöglicht eine Übertragung (Verkauf, Schenkung, Erbe) von Gesellschaftsanteilen anstelle der direkten Übertragung des Vermögenswertes. Hierfür gelten ggf. andere Formvorschriften. Auch lassen sich möglicherweise legale Steuerersparnisse erzielen. Beispiel im deutschen Steuerrecht ist die Übertragung von Anteilen einer grundstücksbesitzenden Kapitalgesellschaft, die keine Grunderwerbsteuer nach sich zieht, wenn Haltefristen (aktuell 10 Jahre) und Beteiligungsgrenzen (aktuell 90 %) nicht verletzt werden (§ 1 Abs. 2b GrEStG).

Allerdings mindert der indirekte Besitz von Vermögen grundsätzlich erst einmal die Transparenz. Es ist nicht auf den ersten Blick ersichtlich, wer „wirtschaftlich Berechtigter“ des Vermögens ist, der über das Vermögen letztlich verfügen kann und dem die Erträge zufließen. Das lässt sich durch Einsicht in weitere Register zwar beheben, aber nur mit Zusatzarbeit und selbstverständlich nur, falls dies möglich ist.[290]

Um Adressen zu finden, die Sitz vieler Unternehmen sind, genügt folgende Abfrage:

```
SELECT c_address, COUNT(c_address)
FROM v_company
GROUP BY c_address
ORDER BY COUNT(c_address) DESC
```

Man erhält erstaunliche Zahlen, hier etwa nur die häufigsten Werte:

[290] Insbesondere eben nicht in vielen ausländischen Jurisdiktionen.

c_address	Anzahl
Mainz_Emy-Roeder-Straße 2, 55129 Mainz.	1697
Pullach i. Isartal_Emil-Riedl-Weg 6, 82049 Pullach i. Isartal.	939
München_Leopoldstr.	458
Grünwald_Südliche Münchner Str.	391
Hamburg_Palmaille 67, 22767 Hamburg.	365
Grünwald_Tölzer Straße 15, 82031 Grünwald.	348
München_Landsberger Str.	346
Hamburg_Bleichenbrücke 10, 20354 Hamburg.	336
München_Maximilianstr.	320
Düsseldorf_Königsallee 106, 40215 Düsseldorf.	316
Grünwald_Bavariafilmplatz 7, 82031 Grünwald.	309
Grünwald_Nördliche Münchner Str.	306
Berlin_Charlottenstraße 4, 10969 Berlin	301
Pullach i.Isartal_Emil-Riedl-Weg 6, 82049 Pullach i.Isartal.	296
Hamburg_Elbchaussee 370, 22609 Hamburg.	274
Bremen_Stephanitorsbollwerk 3, 28217 Bremen	268
... usw.	

Hierzu ist freilich Folgendes anzumerken:

Die Einträge im Handelsregister sind teilweise unpräzise. So ist etwa „München_Leopoldstr.“ der dritthäufigste Wert. Da keine Hausnummer enthalten ist, fallen hierunter aber alle Unternehmen aus dieser Straße, welche keine Hausnummer angegeben hatten.

Die Adresse „Pullach i. Isartal_Emil-Riedl-Weg 6, 82049 Pullach i. Isartal.“ kommt scheinbar doppelt vor, einmal an zweiter Stelle (939) und noch weiter unten (296). Erst ein genauer Blick offenbart, dass sich beide Schreibweisen um ein vorhandenes bzw. fehlendes Leerzeichen unterscheiden. Das verdeutlicht ein generelles Problem, was aber ggf. mit mehr Aufwand lösbar wäre.

Selbstverständlich handelt es sich bei manchen Adressen um große Bürogebäude, die Platz für viele Unternehmen bieten. Das ist natürlich nicht immer so.

An 141 „Adressen“ (beachte: für Fälle, in denen keine Hausnummer eingetragen war, ist dies nur die Straße!) sind mindestens 100 Unternehmen gemeldet:

```
SELECT c_address, COUNT(c_address)
FROM v_company
GROUP BY c_address
HAVING COUNT(c_address) > 99
ORDER BY COUNT(c_address) DESC
```

Für 82.225 Adressen gibt es mindestens 10 Unternehmen derselben Anschrift.

Um Fälle ähnlicher Adressen, sei es durch unterschiedliche Schreibweisen oder teilweise fehlende Hausnummern, zu erkennen, kann man sich die Liste auch alphabetisch sortiert ausgeben lassen:

```
SELECT c_address, COUNT(c_address)
FROM v_company
GROUP BY c_address
HAVING COUNT(c_address) > 99
ORDER BY c_address
```

Hier einige Beispiele:

c_address	Anzahl
Pullach i. Isartal_Emil-Riedl-Weg 6, 82049 Pullach i. Isartal.	939
Pullach i.Isartal_Emil-Riedl-Weg 6, 82049 Pullach i.Isartal.	296
...	
Düsseldorf_Mercedesstr.	118
Düsseldorf_Mercedesstraße 6, 40470 Düsseldorf.	221
...	
Hamburg_Große Elbstr.	179
Hamburg_Große Elbstraße 61, 22767 Hamburg.	143
...	
Eschborn_Mergenthalerallee 10 - 12, 65760 Eschborn.	19
Eschborn_Mergenthalerallee 10-12, 65760 Eschborn.	138

Die Auflistung aller Unternehmen, die unter einer bestimmten Anschrift registriert sind, gelingt mit einer Abfrage wie beispielsweise:

```
SELECT c_name
FROM v_company
WHERE c_address = 'Eschborn_Mergenthalerallee 10-12, 65760 Esch-
born.'
```

Als Resultat erhalten wir:

c_name
BRND X Sales Solutions GmbH
FPI SW Verwaltungs GmbH
SCD Software Center Deutschland GmbH
Infotexx UG (haftungsbeschränkt)
Mavea Yachts Management GmbH
DSP Immobilienmanager UG (haftungsbeschränkt)
Excelsior GmbH
Kleyer Beteiligungsgesellschaft mbH
WEM Experts GmbH
Scapa Holding GmbH
Hanseatic Venture AG
... usw.

Spannend erscheint die Frage, ob alle oder viele der in einem Gebäude ansässigen Firmen auch denselben Officer haben. Für den Gebäudekomplex 'Eschborn_Mergenthalerallee 10-12, 65760 Eschborn.' ist das grundsätzlich nicht der Fall. Nur manchmal ist derselbe Officer bei mehreren Unternehmen eingetragen. Den Firmennamen erhält man über einen JOIN-Befehl, der die Daten beider Tabellen zusammenführt:

```
SELECT DISTINCT c_name, o_name
FROM v_company
LEFT JOIN v_officer ON
v_company.c_number = v_officer.c_number
WHERE v_company.c_address = 'Eschborn_Mergenthalerallee 10-12,
65760 Eschborn.'
ORDER BY o_name
```

c_name	o_name
ALFOUZAN Germany Verwaltungs-GmbH	Abdulmohsin Y Al Johaimy
ALSABIQ Germany Verwaltungs-GmbH	Abdulmohsin Y Al Johaimy
ALSABIQ Eschborn GmbH	Abdulmohsin Y Al Johaimy
Excelsior GmbH	Alexander Bauer
NFL Properties Europe GmbH	Alexander Müller
DIJWS Dienstleistungen GmbH	Alexander Richter
DIJWS Dienstleistungen GmbH	Alexander Schleicher
MI Vertriebs GmbH	Alexander Walleczek
ATBO Dienstleistungsgesellschaft mbH	András Zsolt Böröczky
TRE Invest GmbH	Aneta Tartsch
TRE TRACTION REAL ESTATE GMBH	Aneta Tartsch
MST GmbH	Angelika Hundt
Learn To Trade GmbH	Ann Marie Doctor Agius
Anna Lepper UG (haftungsbeschränkt)	Anna Lepper
...	

Für die Anschrift 'Pullach i. Isartal_Emil-Riedl-Weg 6, 82049 Pullach i. Isartal.' wird beispielsweise Herr Christian Floth als Geschäftsführer von sieben Unternehmen genannt. Eine Internetsuche nach diesem Namen führt auf die Floth Real Estate GmbH,[291] deren Geschäftszweck die Akquirierung und Entwicklung von Immobilien ist. Christian Floth ist der Geschäftsführer und Gründer der Floth Real Estate GmbH.[292] Aus dem Unternehmenszweck kann man ableiten, dass die Existenz von mehrfachen Zweckgesellschaften hier keine Besonderheit darstellt und per se keinen Hinweis für besondere steuerliche Risiken gibt.

In eine andere Richtung geht die Frage, welche Personen auffällig häufig als „officer“ in Erscheinung treten. Näheres erfährt man durch den folgenden Befehl:

[291] https://floth-realestate.com/.

[292] Vgl. https://floth-realestate.com/christian-floth/.

```
SELECT o_name, o_position, o_type, COUNT(c_number)
FROM v_officer
WHERE o_type = 'person'
GROUP BY o_name
ORDER BY COUNT(c_number) DESC
```

Das sind verblüffende Größenordnungen:

o_name	o_position	o_type	Anzahl
Katja Gogalla	Geschäftsführer	person	4124
Antje Borcherdt	Geschäftsführer	person	2069
Angelika Hundt	Geschäftsführer	person	1935
Andreas Koglin	Geschäftsführer	person	1328
Christian Goldbrunner	Geschäftsführer	person	1279
Kerstin Zander	Geschäftsführer	person	1232
Julia Vieth	Geschäftsführer	person	1177
Nicole Lotz	Geschäftsführer	person	1118
Heinz Günter Höhne	Geschäftsführer	person	1117
Cornelia Wendt	Geschäftsführer	person	989
Steffen Kurt Holderer	Liquidator	person	980
Nicole geborene Liebera Lotz	Geschäftsführer	person	892
Thomas Doctor Naumann	Liquidator	person	891
Carsten Eckert	Liquidator	person	864
Achim Bönninghaus	Geschäftsführer	person	824
... usw.			

Grundsätzlich man muss sich schon die Frage stellen, wie man realistischer Weise eine Geschäftsführungstätigkeit bei so vielen Unternehmen einnehmen kann. Allerdings lösen sich manche Fälle bei näherem Hinsehen plausibel auf. Spitzenreiterin ist Frau Katja Gogalla mit sage und schreibe 4.124 Geschäftsführungspositionen.

Eine weitere Abfrage ergibt:

```
SELECT *
FROM v_officer
WHERE o_name = 'Katja Gogalla'
ORDER BY o_city
```

Die Ergebnistabelle deutet mit vier verschiedenen Städten an, dass es sich zunächst einmal vermutlich um vier verschiedene Personen handelt:

o_id	o_name	o_position	o_type	c_number	o_city
2085031	Katja Gogalla	Prokurist	person	R2402_HRB7401	Dortmund
2830440	Katja Gogalla	Geschäftsführer	person	R2402_HRB21749	Dortmund
3374822	Katja Gogalla	Geschäftsführer	person	M1201_HRB80658	Frankfurt am Main
4503612	Katja Gogalla	Geschäftsführer	person	M1201_HRB80659	Frankfurt am Main
3710508	Katja Gogalla	Geschäftsführer	person	K1101R_HRB149869	Hamburg
6639	Katja Gogalla	Geschäftsführer	person	M1201_HRB89553	München
7666	Katja Gogalla	Geschäftsführer	person	M1201_HRB87991	München
8054	Katja Gogalla	Geschäftsführer	person	M1201_HRB94131	München
... usw.					alle München

Dominierend bleibt allerdings Katja Gogalla aus München.

Eine Internetrecherche führt zu folgendem Treffer:[293]

"*Katja Gogalla*

Senior Beraterin

Frau Gogalla ist seit dem Jahr 2004 bei der Blitzstart tätig. Sie ist Industrie- und Handelskauffrau (IHK) und verfügt über mehr als 20 Jahre Erfahrung im Vertrieb."

Die Dame arbeitet also bei der Blitzstart Holding AG aus München, die kommerziell sog. Vorratsgesellschaften als Alternative zur eigenen Unternehmensgründung anbietet.[294] Damit wird in diesem Fall die extrem hohe Anzahl von Firmen verständlich, denn die Firmen sind nicht wirklich werbend tätig. Vielmehr „schlafen" sie eher vor sich hin und warten auf einen passenden Erwerber.

Er versteht sich von selbst, dass solche plausiblen Erklärungen nicht immer vorliegen. In solchen Fällen können Indizien für besondere steuerliche Risiken gegeben sein.

Besonders aufschlussreich könnten solche Auswertungen sein, wenn sie sich nicht auf das deutsche Unternehmensregister beschränken, sondern auch soweit möglich entsprechende **Register anderer Staaten** einbeziehen. Bereits innerhalb der EU gibt es ja mehrere Staaten, die nach weit verbreiteter Einschätzung als „Steueroasen" gelten (ausführlich dazu Kapitel II., Abschnitt C. 3. c) Standort und Rechtsform auf Seite 57). Hierzu zählen u.a. Luxemburg, Niederlande, Irland, Malta und Zypern.

[293] https://www.blitzstart.com/de/ueber-uns/team/.

[294] Vgl. https://www.blitzstart.com/de/.

Beispiele für Industrieländer (also keine reinen Steueroasen) außerhalb der EU sind USA, Schweiz, Hongkong und das United Kingdom.

Es liegt auf der Hand, dass die Erschließung ausländischer Datenquellen die Möglichkeiten vervielfältigt. Insbesondere wäre auch ein **Abgleich des deutschen Transparenzregisters** mit den Inhalten ausländischer Register möglich, wenn der unmittelbare Anteilseigner in einem ausländischen Unternehmen besteht.

4.2.2 Linked Leaks und FactForge

Linked Leaks ist eine Datensammlung, die – wie der Name schon andeutet – auf den Enthüllungen zu den Panama Papers aufbaut. Letztere wurden ja vom International Consortium of Investigative Journalists (ICIJ) veröffentlicht (siehe Abschnitt III. B. 5. Steuerleaks auf den Seiten 89ff.).

Die Firma Ontotext hat die öffentlichen Daten der Panama Papers zu Demonstrationszwecken in ihr Graphen-DBMS „GraphDB“ eingelesen.[295] Diese Daten wurden jedoch noch um weitere Daten angereicht. Dazu zählen geographische Angaben aus DBPedia und aus GeoNames. Bereits hiermit werden umfangreichere Analysen möglich, da beispielsweise in den Panama Papers erwähnte Städte automatisch ihrem Land zugeordnet werden können.

Linked Leak lässt sich über eine Webseite als SPARQL Endpoint abfragen.[296] Alternativ ist auch der Download des kompletten Datensatzes als zip-Datei möglich. Das DBMS „GraphDB Free“ wird ebenfalls kostenlos zur Verfügung gestellt:[297]

Es handelt sich somit um Linked Open Data:

“*Linked Leaks, an Ontotext portal, publishes this data as a knowledge graph, according to the Linked Open Data principles. This allows one to enter the URL identifier of the an entity or a person (say, http://data.ontotext.com/resource/leaks/entity-123456) in a web browser to see all the information available in for this entity in the database. Applications can retrieve the relevant information for a re-source making HTTP GET with its URL.*

[295] Vgl. hierzu und zum Folgenden http://data.ontotext.com/linkedleaks.

[296] http://data.ontotext.com/linkedleaks.

[297] ftp://ftp.ontotext.com/pub/leaks/rdf/rdf.zip.

Data allows for all sorts of discovery and analytics queries such as:

- *Companies that have more than one shareholder in common with a given one;*
- *Companies related to a given shareholder (be it person or organization), including control relationships;*
- *Companies that control other companies in the same country, through company in an off-shore zone;*
- *Most popular offshore jurisdictions*."[298]

Die Webseite verweist dazu auf einige Beispielabfragen.[299]

"LL2: Country pairs by ownership statistics" zeigt für die Panama Papers, welche Länder am häufigsten über Eigentumsverhältnisse verbunden sind. Die SPARQL-Abfrage dazu lautet:[300]

```
PREFIX onto: <http://www.ontotext.com/>
PREFIX leak: <http://data.ontotext.com/resource/leak/>
PREFIX leaks: <http://data.ontotext.com/resource/leaks/>
SELECT ?owner_country ?owner_country_name ?entity_country ?en-
tity_country_name (COUNT(*) as ?count)
FROM onto:disable-sameAs
{
    ?owner leak:OWNER_TRANSITIVE ?entity .
    ?owner leak:hasCountry ?owner_country .
    FILTER(?owner_country != leaks:country-XXX)
    ?entity leak:hasCountry ?entity_country .
    FILTER(?entity_country NOT IN (leaks:country-XXX, ?owner_coun-
try ))
    ?owner_country leak:name ?owner_country_name .
    ?entity_country leak:name ?entity_country_name .
}
GROUP BY ?owner_country ?owner_country_name ?entity_country ?en-
tity_country_name
ORDER BY DESC(?count)
```

[298] http://data.ontotext.com/linkedleaks.

[299] Siehe auch Code und Erläuterungen auf Github. https://github.com/Ontotext-AD/leaks.

[300] http://data.ontotext.com/sparql?savedQuery-Name=LL2:%20Country%20pairs%20by%20ownership%20statistics&execute=true.

Für Deutschland erhält man folgende Ergebnisstabelle:

Land 1	Name 1	Land 2	Name 2	Anzahl
leaks:country-DEU	Germany	leaks:country-VGB	Virgin Islands, British	"1762"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-SGP	Singapore	"282"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-IDN	Indonesia	"175"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-HKG	Hong Kong	"109"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-IND	India	"73"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-MYS	Malaysia	"73"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-CHE	Switzerland	"46"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-THA	Thailand	"38"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-CYM	Cayman Islands	"35"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-LUX	Luxembourg	"31"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-RUS	Russian Federation	"24"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-TWN	Taiwan, Province of China	"22"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-GBR	United Kingdom	"19"^^xsd:integer
leaks:country-DEU	Germany	leaks:country-ITA	Italy	"19"^^xsd:integer

Die einzelnen Treffer, die hinter der aggregierten Anzahl stecken, ließen sich selbstverständlich auch auflisten.

Ein Webinar mit dem Titel „Diving in Panama Papers and Open Data to Discover Emerging News" stellt verschiedene Auswertungsmöglichkeiten dar.[301] Es wurde aufgezeichnet und ist frei auf Youtube verfügbar.[302]

Ein weiteres Demonstrationsbeispiel von Ontotext neben LinkedLeaks ist **FactForge** (siehe auch Abschnitt III. B. 7. auf Seite 95).[303] Hier werden noch umfangreichere Datenquellen integriert wie GLEI, WorldFacts,[304] WordNet[305] und Now News.[306]

[301] https://www.ontotext.com/knowledgehub/webinars/discover-emerging-news-with-open-data/?utm_source=LinkedLeaks&utm_medium=refferal&utm_campaign=LinkedLeaks.

[302] https://www.youtube.com/watch?v=wKLkAAafAdY.

[303] https://www.ontotext.com/blog/exploring-linked-open-data-factforge/.

[304] http://worldfacts.us/.

[305] https://wordnet.princeton.edu/ für Synonyme.

[306] http://now.ontotext.com für aktuelle Nachrichten (laufend aktualisiert).

Die dahinter stehenden Ontologien sind sehr umfangreich und extrem komplex. [307] Auch hier gibt es einen SPARQL Endpoint.[308]

Die SPARQL-Abfrage für das Beispiel "# F05: Suspicious control chain through offshore company" lautet etwa:

```
PREFIX onto: <http://www.ontotext.com/>
PREFIX fibo-fnd-rel-rel: <http://www.omg.org/spec/EDMC-
FIBO/FND/Relations/Relations/>
PREFIX ff-map: <http://factforge.net/ff2016-mapping/>
SELECT *
FROM onto:disable-sameAs
WHERE {
    ?c1 fibo-fnd-rel-rel:controls ?c2 .
    ?c2 fibo-fnd-rel-rel:controls ?c3 .
    ?c1 ff-map:primaryCountry ?c1_country .
    ?c2 ff-map:primaryCountry ?c2_country .
    ?c3 ff-map:primaryCountry ?c1_country .
    FILTER (?c1_country != ?c2_country)
    ?c2_country ff-map:hasOffshoreProvisions true .
}
```

Hier werden Einflussketten aufgelistet, bei denen ein Unternehmen (quasi der Enkel) indirekt von einem anderen Unternehmen (quasi die Mutter) beherrscht wird, wobei ein weiteres Unternehmen (Tochter) mit Sitz in einer Steueroase zwischengeschaltet ist.

[307] Siehe http://factforge.net/relationships und http://factforge.net/hierarchy.

[308] http://factforge.net/sparql.

4.2.3 Projekt execGraph

Das dritte Beispiel stammt mit der Wirtschaftsprüfung aus einem Gebiet, das nur indirekt mit der steuerlichen Betriebsprüfung verbunden ist. Gemeinsamkeit ist jedoch die Frage, ob Indikatoren für besondere, erhöhte Risiken vorliegen.

Konkreter aktuelle Bezug ist die Wirecard AG. Angesichts der Unregelmäßigkeiten der Rechnungslegung und der mangelnden Aufsicht über das deutsche Finanzdienstleistungsunternehmen Wirecard AG, die schließlich in der Insolvenz und der Verhaftung des Vorstandsvorsitzenden Markus Braun endeten, wurden die Forderungen nach einer umfassenderen Finanzaufsicht lauter.[309] Journalisten und Wissenschaftler waren erstaunt, wie ein Unternehmen, das im Deutschen Aktienindex (DAX) unter den 30 größten börsennotierten Unternehmen Deutschlands gelistet ist, Aktionäre, Finanzaufsichtsbehörden und die Öffentlichkeit über einen so langen Zeitraum täuschen konnte.[310]

Das Forschungsprojekt „execGraph“ wurde begonnen, um Risikohinweise in zwei Richtungen zu gewinnen:

Um als Abschlussprüfer tätig zu sein, darf keine Besorgnis der Befangenheit vorliegen (§ 328 Abs. 2 HGB), welche die Unabhängigkeit des Wirtschaftsprüfers bzw. der Wirtschaftsprüfungsgesellschaft gefährden könnte. Hierzu hat der Gesetzgeber umfangreiche Regelungen explizit kodifiziert (§ 318 Abs. 3 und 4 HGB, § 319a HGB, § 319b HGB), welche allerdings von Zeit zu Zeit verändert – im Sinne von zunehmend verschärft – wurden. Generell könnten also auch „Beziehungen“ schwächerer Art graduell die Unabhängigkeit schwächen. Hinzu kommt, dass die angewandten Kriterien sich in den allermeisten Fällen nur auf die Gegenwart beziehen.[311] Frühere Ausschlussgründe, die heute nicht mehr bestehen, sind rechtlich unbeachtlich, auch wenn sie möglicherweise noch nachwirken. Die Existenz von schwächeren und/oder vergangenen Beziehungen zwischen dem Abschlussprüfer und den Organen – Vorstände, Aufsichtsräte – des zu prüfenden Unternehmens sollen als erstes im ExecGraph erfasst werden.

[309] Vgl. Krahnen / Langenbucher (2020); Wortham / Liebscher / Christian (2020); Véron (2020).

[310] Vgl. Alderman / Schuetze (2020); Bartz et al. (2020).

[311] Ausnahme ist etwa die Regelung nach § 319 Abs. 3 Nr. 5 HGB (hoher Honoraranteil von dem Mandanten innerhalb der letzten 5 Jahre).

Daneben spielt auch die Qualifikation des Aufsichtsrates in seiner Gesamtheit (§ 107 AktG, § 100 Abs. 5, 2. HS), des Prüfungsausschusses bei kapitalmarktorientierten Unternehmen (§ 324 Abs. 1 HGB) sowie des sog. Finanzexperten im Aufsichtsrat (§ 100 Abs. 5, 1. HS) eine wichtige Rolle. Auch hierfür gibt es kodifizierte Mindestanforderungen, die jedoch nicht alleine Maßstab sein können. Der ExecGraph soll daher auch das Qualifikationsgefüge im Aufsichtsrat so gut wie möglich mehrdimensional – Kompentenzportfolio, Erfahrungswissen – abbilden.

Dem Beispiel liegt die Annahme zugrunde, dass Risiken in Hinblick auf die Korrektheit der externen Rechnungslegung auch von Bedeutung für steuerliche Risiken sein können.

Die Aufsicht von Unternehmen steht somit im Mittelpunkt des execGraph. Ziel ist es, Möglichkeiten zur Verbesserung der Rechenschaftspflicht und Transparenz im Bereich der Regulierung zu erkunden. Fraglich ist, inwiefern Mitglieder des Vorstands und Aufsichtsrats miteinander verknüpft sind. Verknüpfungen können bspw. aufgrund gemeinsamer Tätigkeiten, Mitgliedschaften in denselben Organisationen oder auch gemeinsamer Ausbildung bestehen.

Eine Möglichkeit, diese Informationen zu strukturieren, ist ein Knowledge Graph. Die Knoten stellen reale Entitäten und Ressourcen dar, während die Kanten ihre Beziehung zueinander beschreiben.[312] Die Stärke eines Wissensgraphen liegt in seiner Fähigkeit, die begrenzten Informationen in der Datenbank durch Links zu externen, im Internet verfügbaren offenen Graphen-Datenbanken anzureichern. Da Wissensgraphen die Beziehungen zwischen Entitäten betonen, ermöglichen sie die Analyse und Visualisierung dieser Beziehungen.

Das Gesamtziel des Projekts ist die Erforschung des Einsatzes von Wissensgraphen als Instrument zur Untersuchung der Beziehungen zwischen Vorständen, Aufsichtsräten und Wirtschaftsprüfern in den deutschen DAX-30-Unternehmen. In diesem Sinne gibt es fünf Teilziele: (1) Datenbankgenerierung (2) Graphenmodellierung (3) Graphengenerierung (4) Graphenanalyse und (5) Graphenvisualisierung.

[312] Vgl. Ehrlinger / Wöß (2016).

Die *Datenbankgenerierung* umfasst das Bereinigen vorhandener Daten, das Extrahieren neuer Daten aus verschiedenen Quellen und das Erstellen einer zentralen Datenbank, die den üblichen Datenbanknormalisierungsregeln entspricht und als Grundlage für den Wissensgraphen dient. Die *Graphenmodellierung* umfasst den Entwurf eines Modells auf der Grundlage der verfügbaren Daten, das eine sinnvolle Durchquerung und Analyse des Graphen ermöglicht. Die *Graphengenerierung* beschreibt eine Instanziierung des Graphenmodells. Die *Graphenanalyse* beschreibt alle Aufgaben, die mit der Suche nach aussagekräftigen Erkenntnissen und der Evaluierung des Projekts selbst zusammenhängen. Die *Graphvisualisierung* zielt darauf ab, eine visuelle Darstellung des Graphen zu liefern, die Interaktion und Navigation ermöglicht. Das Ziel ist die Verwendung von Programmen (VisiNav und Tarsier), die eine grafische Benutzeroberfläche (GUI) zur Verfügung stellen und somit die Einstiegshürden für Benutzer mit weniger technischem Know-how beseitigen.

a) Speicherung der Daten

Daten in einem Excel- oder csv.-Format sind für den Knowledge Graph nicht geeignet. Dementsprechend müssen die Daten erst normalisiert werden. Der Schwerpunkt liegt dabei auf der Benutzerfreundlichkeit für Benutzer ohne technisches Fachwissen, indem MS Excel als Speicherformat beibehalten wird, während gleichzeitig versucht wird, die in relationalen Datenbanken üblichen Prinzipien zu befolgen. Folglich wird eine einzige MS Excel-Datei erstellt, die alle erforderlichen Informationen in 11 Registerkarten enthält, wie in Tabelle 2 zu sehen ist. Allen Registerkarten, die sich auf Ressourcen beziehen (in Tabelle 2 blau hervorgehoben), wird eine eindeutige ID zugewiesen, die bei der Erstellung des Diagramms in eine URI umgewandelt wird, sowie gegebenenfalls Links zu externen Datenbanken (z. B. *wikidata_sameAs*).

Registerkarte	Beschreibung
Person	Weist jeder Person *(personEntity)* eine eindeutige ID zu. Enthält persönliche Informationen zu jeder Person im Diagramm (z. B. Name, Geburtsjahr).
Unternehmen	Weist jedem Unternehmen eine eindeutige ID zu *(companyEntity)*. Enthält Informationen zu jedem Unternehmen im Diagramm (z. B. Name, ISIN, DAX-Mitgliedschaft).
Universität	Weist jeder Hochschule eine eindeutige ID zu *(universityEntity)*. Enthält Informationen zu jeder Hochschule im Diagramm (z. B. Name, Link zu anderen Datenbanken).
Städte	Weist jeder Stadt eine eindeutige ID zu *(cityEntity)*. Enthält Informationen zu jeder Stadt im Diagramm (z. B. Name, Link zu anderen Datenbanken).
Länder	Weist jedem Land eine eindeutige ID zu *(countryEntity)*. Enthält Informationen zu jedem Land im Diagramm (z. B. Name, Link zu anderen Datenbanken).
Position	Enthält Informationen über Mitarbeiterpositionen, die eine Person in ihrer Karriere in einem Unternehmen innehatte oder derzeit innehat. Dient als Bindeglied zwischen einem Unternehmen und einer Person.
Bildung	Enthält Informationen über Bildungsprogramme oder Abschlüsse, die eine Person erworben hat. Dient als Bindeglied zwischen einer Universität und einer Person.
Prüfung	Enthält Informationen über Prüfungen, die ein Unternehmen durchlaufen hat. Dient als Bindeglied zwischen einer Wirtschaftsprüfungsgesellschaft, dem geprüften Unternehmen und den Wirtschaftsprüfern (einer Person).
Geschäfte	Enthält Informationen über Eigengeschäfte, die eine Person mit einem Unternehmen getätigt hat. Dient als Verbindung zwischen einem Unternehmen und einer Person.
Beziehungen	Enthält Informationen über Beziehungen zwischen Personen (z. B. Ehepartner, Geschwister). Dient als Verbindung zwischen zwei Personen.
Organisation	*Wurde aufgrund der geringen Datenmenge eingestellt: Enthält Informationen über die Mitgliedschaft einer Person in Organisationen. Dient als Bindeglied zwischen einer Person und einer Organisation.*

Abb. 15: Übersicht der Registerkarten der Datenbank Herkunft der Daten

Die Daten wurden aus einer Vielzahl von Quellen extrahiert, wobei die meisten Daten von den DAX-Unternehmen selbst zur Verfügung gestellt wurden, um Transparenz für ihre Aktionäre zu schaffen. Allerdings waren die Lebensläufe teilweise direkt auf der Webseite oder im PDF-Format sowie strukturiert (Liste aller Positionen) oder unstrukturiert (reine Textbeschreibung) enthalten.

Daher ist es eine Herausforderung, Daten zu scrapen oder den Prozess zu automatisieren. Zusätzlich werden offene Graph-Datenbanken wie DBpedia[313] und WikiData[314] genutzt, allerdings waren die Informationen hier nur unvollständig oder veraltet. Außerdem werden Daten aus einer Bloomberg-Datenbank entnommen. Eine weitere wertvolle Datenquelle ist LinkedIn, wo die Personen selbst Informationen bereitstellten und pflegen.

b) Umfang der Daten

Die Datenbasis besteht aus 745 Personen darunter Führungskräfte, Aufsichtsratsmitglieder und Wirtschaftsprüfer. Mit Hilfe von OpenRefine[315] kann etwa ein Drittel (32 %) der Personen mit Wikidata verknüpft werden. Darüber hinaus finden sich 1.203 Unternehmen in der Datenbank. Hier kann mehr als die Hälfte (53 %) mit Wikidata verknüpft werden. Insgesamt umfasst unsere Datenbank 5.116 Stellen, von denen 1.830 im Dezember 2020 besetzt waren, sowie 1.128 Bildungsabschlüsse (zu den Formen siehe Abbildung 16 und Studiengänge/Bildungskategorien (siehe Abbildung 17) an 341 Hochschulen, die von den gesuchten Personen absolviert wurden.

Grad
Hauptschule
Mittlere Reife
Abitur
Ausbildung
Bachelor
Diplom
MBA
Meister
Staatsexamen
PhD
Weiterbildung

Abb. 16: Abschlüsse

[313] https://wiki.dbpedia.org/.
[314] https://www.wikidata.org/wiki/Wikidata:Main_Page.
[315] https://openrefine.org/.

Feld 1/ Feld 3	Feld 2/ Feld 4
Wirtschaftswissenschaften	VWL BWL FACT Betriebswirtschaftslehre Kommunikationswissenschaft Marketing Wirtschaft
Ingenieurswissenschaften	Elektrotechnik Informatik Maschinenbau Maschinenbau Luft- und Raumfahrttechnik Agrartechnik
Naturwissenschaften	Biologie Chemie Physik Mathematik Medizin
Geisteswissenschaften	Theologie Philosophie Journalismus Politikwissenschaften Soziologie Geschichte Psychologie Anglistik Kunst Sport
Rechtswissenschaften	
Pädagogik	
Linguistik	Englisch, etc.
Andere	

Abb. 17: Studiengänge/Bildungskategorien

c) Verknüpfung der Daten

Während die Verknüpfung von Entitäten im Graphen mit Datenbanken wie WikiData oder DBpedia manuell erfolgen kann, ist OpenRefine ein nützliches Werkzeug, um diesen Schritt zu automatisieren. Nachdem eine Excel- oder csv-Datei geöffnet wurde, kann ein Projekt erstellt werden. OpenRefine gleicht dann die Namensstrings in ihrer Datei mit den Namen der Entitäten in der ausgewählten Datenbank ab.

d) Graph Modellierung

Die Erstellung des Graphen erfolgte mit tarql[316] , einem in Java geschriebenen Kommandozeilen-Tool zur Konvertierung von CSV-Dateien in RDF mit SPARQL-Syntax. Tarql benötigt drei Dinge: (1) einen Befehl, (2) eine csv-Datei und (3) eine SPARQL-Konstruktabfrage.

Der Befehl weist tarql im Wesentlichen an, eine bestimmte SPARQL-Konstruktabfrage für eine bestimmte csv-Datei durchzuführen. Variablen ('?example') werden mit den Spaltennamen in der csv-Datei abgeglichen. Um den Befehl auszuführen, muss zuerst die Kommandozeile im tarql-1.2 Ordner geöffnet und das tarql Unix Executable (siehe Pfad in Beispiel 3) ausgeführt werden.

Die csv-Datei kann für jede Registerkarte in MS Excel erstellt werden, allerdings traten mehrere Probleme bei der Formatierung auf. Es ist wichtig, dass keine Bezeichner (BeispielEntity-Spalten) Sonderzeichen (‚ü', ‚ö', ‚ä', ‚ß' oder Leerzeichen dazwischen) enthalten und dass keine Zelle Doppelpunkte oder Semikolons enthält. Wir haben einen Workaround für die Konvertierung in UTF-8 csv-Dateien gefunden, der bisher gut funktioniert hat. Man erstellt einfach eine leere Google-Tabelle und fügt alle Daten aus einer Registerkarte als Werte ein und speichert die Tabelle als csv-Datei.

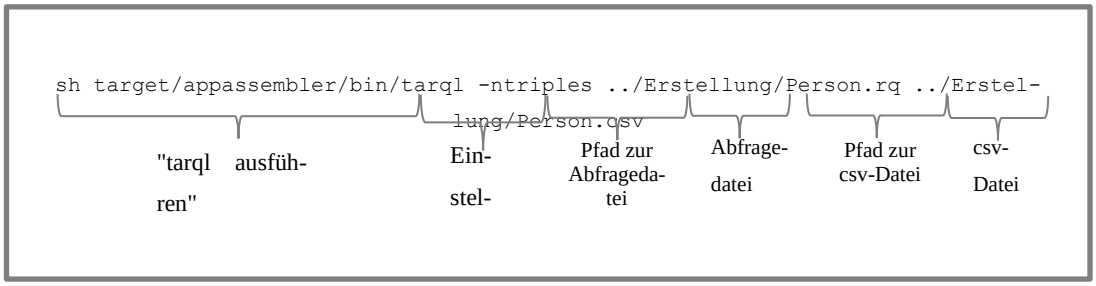


Abb. 18: tarql Befehl

Die SPARQL-Konstruktabfragen können in jedem Texteditor geschrieben und als .rq-Datei gespeichert werden. Zu Beginn kann es hilfreich sein, einen Blick auf etablierte Abfragen zu werfen, um ein Verständnis für Anforderungen und Syntax zu bekommen, wie in Beispiel 4 zu sehen ist.

[316] https://tarql.github.io/.

```
PREFIX ex: <http://execgraph.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX-Schema: <http://schema.org/>
PREFIX dbpprop: <http://dbpedia.org/property/>

KONSTRUKTION {
         ?person ex:hasRole [
                           eine schema:EmployeeRole;
                           schema:roleName ?jobTitle;
                           schema:startDate ?startgYear;
                           schema:endDate ?endgYear;
                           ex:roleWith ?company;
                           ex:inDivision ?divisonName;
                           ex:isCurrent ?currentBool;
                           ex:isSupervisoryBoard ?positionARBool;
                           ex:isManagementBoard ?positionVOBool
                           ]
}
FROM <Datei:Position.csv>
WHERE {
         BIND (URI(CONCAT('http://execgraph.org/', ?positionID)) AS ?position)
         BIND (URI(CONCAT('http://execgraph.org/', ?personEntity)) AS ?person)
         BIND (URI(CONCAT('http://execgraph.org/', ?entityCompany)) AS ?company)
         BIND (xsd:gYear(?startYear) AS ?startgYear)
         BIND (xsd:gYear(?endYear) AS ?endgYear)
         BIND (xsd:boolean(?current) AS ?currentBool)
         BIND (xsd:boolean(?positionAR) AS ?positionARBool)
         BIND (xsd:boolean(?positionVO) AS ?positionVOBool)
}
```

Abb. 19: SPARQL-Konstrukt-Abfrage

Nach dem Ausführen des Befehls werden die Tripel automatisch generiert und können kopiert und eingefügt werden. Durch Einfügen aller Tripel aus allen Registerkarten haben wir dann den fertigen Graphen erstellt. Das Ergebnis wird als Turtle-Datei (.ttl) gespeichert.

e) Graphische Analyse

Die Abfrage birthPlaces_map veranschaulicht die Nutzung offener Graphen-Datenbanken über unsere eigenen Daten hinaus, da sie Stadtkoordinaten aus Wikidata abfragt und uns somit erlaubt, Geburtsorte von DAX-Mitgliedern wie unten in Abbildung 20 gezeigt abzubilden. In Hinblick auf die Corporate Governance kann ein gewisses Ausmaß an geographischer – insbesondere internationaler – Diversität dazu beitragen, einseitig verengte Blickwinkel vermeiden. Sehr eng örtlich konzentrierte Herkunftsorte können auf persönliche Beziehungen hindeuten, die der geforderten Unabhängigkeit zuwiderlaufen. Es wären selbstverständlich Ausschnitte der Karte vergrößerbar. Auch ließen sich aus den Stadtkoordinaten die Entfernungen berechnen, ohne dass eine Karte dargestellt werden muss.

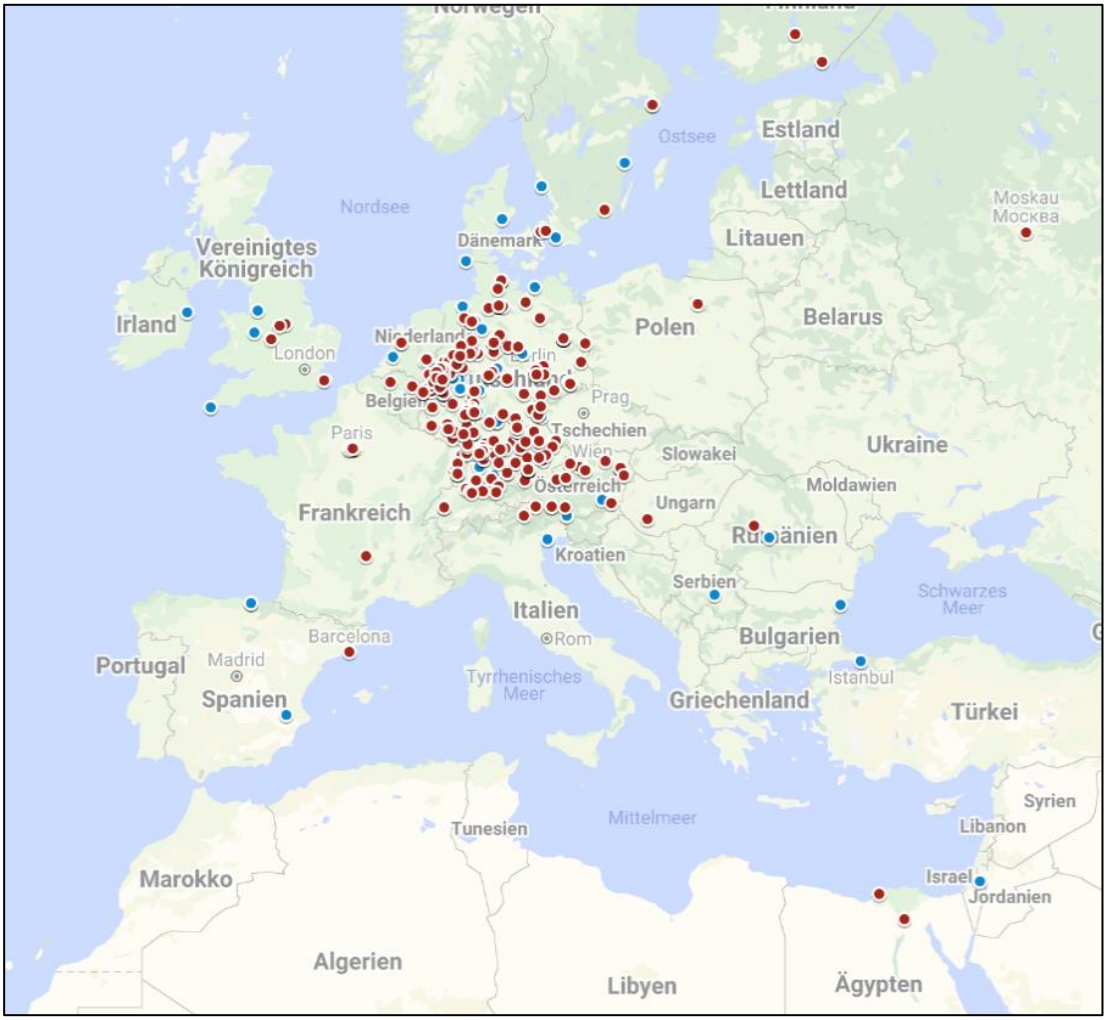


Abb. 20: Karte mit Geburtsorten der DAX-Vorstände (blau) und Aufsichtsräte (rot)

f) Visualisierung

Gephi[317] ist ein intuitives und bewährtes Werkzeug zur Visualisierung von Graphen und bietet eine Vielzahl verschiedener Funktionen zur Netzwerkanalyse. Der Nachteil ist, dass das Tool nicht sehr gut skalierbar ist und Schwierigkeiten bei der Darstellung großer Diagramme hat. Zunächst muss der Turtle-Graph (oder Subgraph) in das .gexf-Format konvertiert werden. Ein Beispiel, wie das geht, finden Sie in den Skripten mit der Endung _to_gexf in execgraph/python_conversion. In Gephi können die .gexf-Dateien importiert und in verschiedenen Formen (z.B. Graphvisualisierung, Datentabelle) und personalisiert (z.B. Farbschema, Knotengröße) dargestellt werden. Darüber hinaus können verschiedene Metriken zur Netzwerkanalyse berechnet werden, wobei die nützlichste wahrscheinlich PageRank ist.

[317] https://gephi.org/.

Die von Page et al. (1999) vorgeschlagene PageRank-Metrik berechnet die wahrgenommene Wichtigkeit eines Knotens in einem gerichteten Graphen und kann durch den folgenden vereinfachten Algorithmus ausgedrückt werden:

$$R(u) = c \sum_{u \in B_u} \frac{R(v)}{N_v}$$

Wenn *u* ein Knoten im Graphen ist, ist *v* ein Knoten in der Menge B_u ist, die alle Knoten enthält, die auf die Seite *u* verlinken, und N_v die Anzahl der Links vom Knoten *v* ist. Der PageRank des Knotens u wird dann iterativ berechnet, indem der PageRank jedes Knotens, der auf den Knoten *u* verlinkt, geteilt durch seine ausgehenden Links summiert wird. Dabei werden Links von Knoten mit höherem PageRank stärker gewichtet als Links von solchen mit niedrigerem PageRank. Der anfängliche PageRank-Wert für jeden Knoten in Gephi wird auf $\frac{1}{n}$ gesetzt, wobei *n* die Anzahl der Knoten im Graphen ist.

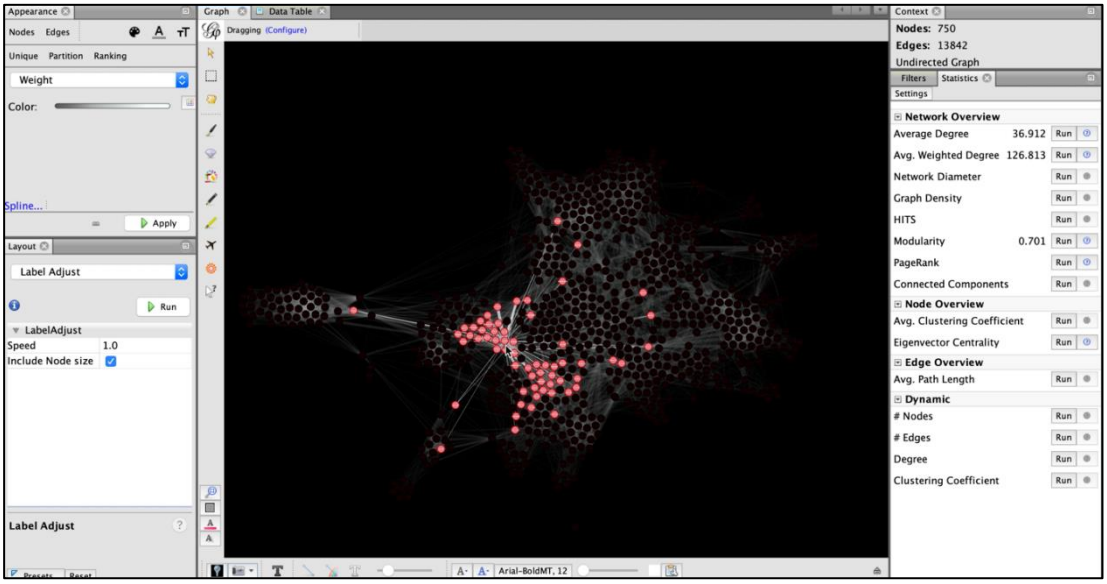


Abb. 21: Screenshot der Gephi-Visualisierung

Graphistry[318] ist ein fortschrittlicheres Tool zur Visualisierung von Graphen, das in Python verwendet werden kann und die beste Leistung der in diesem Abschnitt genannten Tools aufweist.

[318] https://www.graphistry.com/.

Der Nachteil ist, dass das Tool nicht so viele Funktionen zur Netzwerkanalyse bietet wie Gephi. Graphistry bietet allerdings die Möglichkeit einer dynamischen Visualisierung.

Tarsier[319] bietet eine weitere Möglichkeit, RDG Knowledge Graphs in 3D zu visualisieren. Der Vorteil von Tarsier liegt darin, dass verschiedene Ebenen erstellt werden können. In unserem Beispiel etwa für:

- Unternehmen
- Aufsichtsrat
- Vorstand
- Wirtschaftsprüfer.

Auswertungen können auf diese Weise vielfältig vorgenommen werden. Es können bspw. die folgenden Verflechtungen vorliegen:

- Tätigkeit in mehr als einem Gremium (z.B. Aufsichtsrat zeitgleich bei 2 (konkurrierenden) Unternehmen).
- Mitglieder eines Gremiums oder Wirtschaftsprüfer und Aufsichtsratsmitglied studierten zeitgleich an selber Universität.
- Ehemaliger Wirtschaftsprüfer einer Gesellschaft wechselt in Aufsichtsrat oder Vorstand der Gesellschaft.
- Verwandtschaften oder private Freundschaften zwischen Vorstand und Aufsichtsrat bzw. Vorstand und Wirtschaftsprüfer bzw. Aufsichtsrat und Wirtschaftsprüfer oder innerhalb eines Gremiums.
- Erfahrung der Mitglieder in Position in anderer Gesellschaft oder anderem Gremium.
- Aufsichtsrat oder Vorstand war ehemaliger Vorstand oder Aufsichtsrat in der gleichen Gesellschaft.
- Wurden Personen wegen ihrer Eignung/Qualifikation eingestellt oder möglicherweise aufgrund personeller Verflechtung.

[319] Entwickelt von Fabio Viola, Universität Bologna.

- Gleiche Nationalität, Herkunft oder Geburtsort von Aufsichtsräten, Vorständen oder Wirtschaftsprüfern.

Die **Analyse der Wirecard AG** führt zu einigen aufschlussreichen Erkenntnissen.[320] Wirecard hatte im Vergleich zu anderen DAX-Gesellschaften wie beispielsweise der Allianz SE nur eine kleine Anzahl von Verbindungen zu anderen Unternehmen. Es war eine Art „closed shop". Über die Mitglieder von Vorstandes war nur wenig öffentlich bekannt. So enthält die Datenbank zum Wirecard-Vorstand Jan Marsalek nur 48 Triples, zu Oliver Bäte, Vorstandsvorsitzender der Allian, hingegen 123 Triples. Der Aufsichtsrat von Wirecard umfasste nur wenige Personen (anfangs drei, später fünf, dann sechs). Diese hatten keine Aufsichtsratserfahrung in (anderen) DAX30-Unternehmen und ganz überwiegend zeitlich nur sehr kurze Aufsichtsratserfahrung (). Ein Mitglied verließ den Aufsichtsrat bereits nach einem Jahr wieder, was Fragen nach dem Grund hierfür aufwirft.

Die umseitige Abbildung zeigt die allgemeine **Struktur des Graphen**, unabhängig von Wirecard.

Die Datensammlung wird laufend erweitert. Ein **probeweiser Zugriff** (Betaversion) auf Daten des execGraph ist möglich unter: http://tc.ontologycentral.com/execgraph/.

[320] Grümmer (2021).

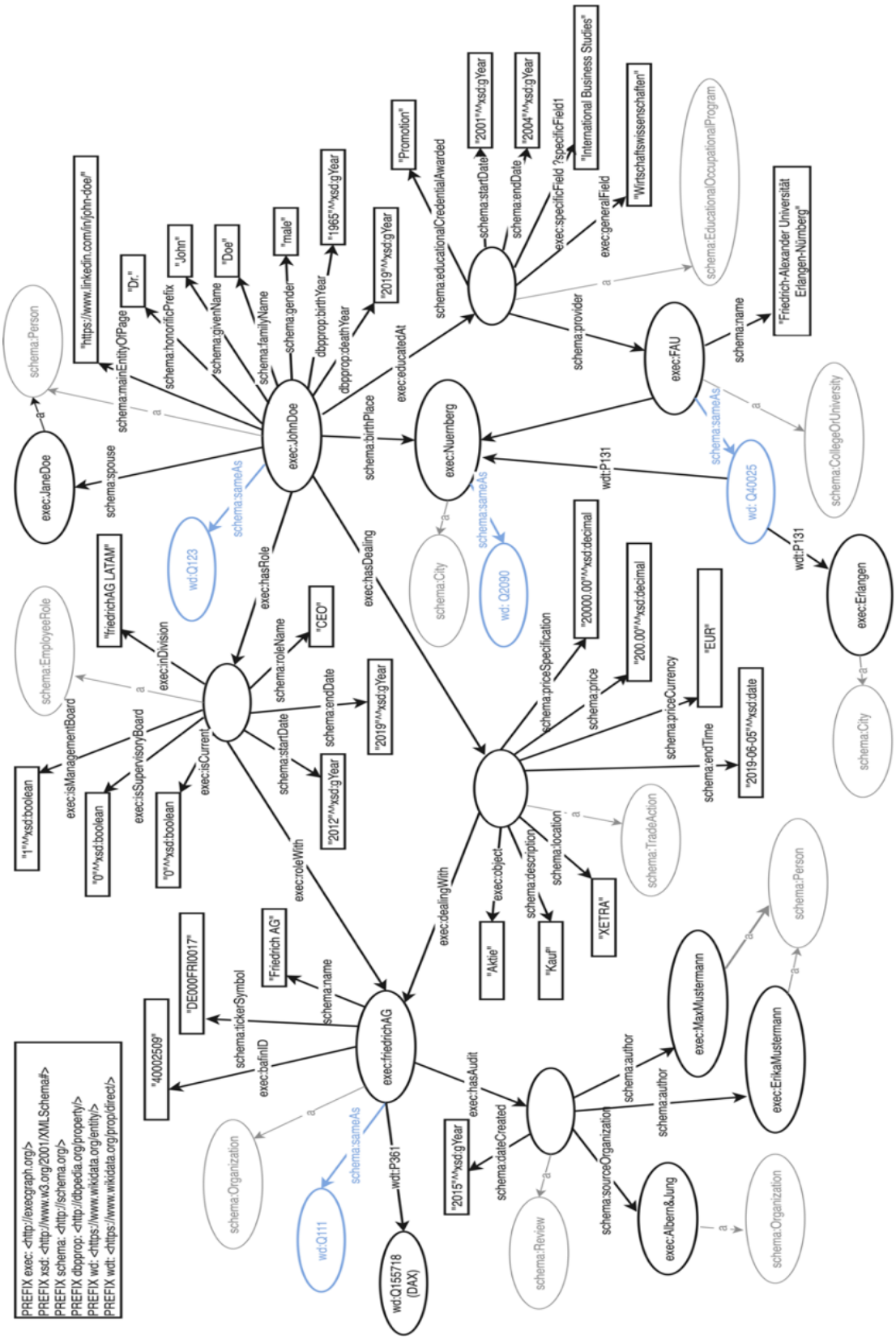


Abb. 22: Struktur des execGraph
Eigene Abbildung

5 Zusammenfassung und Ausblick

Die wesentlichen Erkenntnisse dieser Arbeit lassen sich in den folgenden Thesen zusammenfassen:

(1) Es gibt bereits heute umfangreiche technische Möglichkeiten zur Nutzung von Alternativen Daten, die bei weitem noch nicht ausgeschöpft werden. Da der Rechtsrahmen langfristig als veränderbar angesehen wird, blendet die vorliegende Untersuchung juristische Fragen aus. Augenscheinliche Einsatzmöglichkeiten bestehen einerseits in Konsistenzprüfungen (der bislang intern über die Steuerpflichtigen vorliegenden Daten mit externen Informationen aus den Alternativen Daten), andererseits in der Identifikation von Risikobereichen, bei denen „Warnsignale" (Red Flags) für versehentlich oder gar absichtlich falsche bzw. fehlende Steuererklärungen vorliegen.

(2) Für Konsistenzprüfungen (etwa von CbCR Berichten) lassen sich außerhalb der Finanzverwaltung keine konkreten Auswertungen zeigen, da die Daten dem Steuergeheimnis unterliegen. Jedoch illustrieren drei beispielhafte Auswertungen das Potenzial für die Gewinnung von „Warnsignalen" aus Alternativen Daten.

(3) Die Beispiele verdeutlichen jedoch auch, dass es sich häufig um „unscharfe" Daten handelt. Probleme für die Auswertung entstehen aufgrund von lückenhaften Angaben (z.B. nur die Straße ohne Hausnummer als Adresse), verschiedenen Schreibweisen (z.B. „Manfred Mustermann", „M. Mustermann", „Dr. M. Mustermann", „Manfred G. Mustermann", Mustermann Manfred", „Manni Mustermann"), Tippfehlern („Manferd Mustermann"), Namensgleichheit unterschiedlicher Personen (viele Datensätze haben keinen eindeutigen Identifier/Primärschlüssel) und vieles andere.

(4) Die Verknüpfung mehrerer verschiedener Datensätze wird ebenfalls dadurch behindert, dass – selbst wenn sie jeweils einen eindeutigen Identifier besitzen – diese Identifier/Primärschlüssel nicht dieselben sind. Eine wechselseitige Zuordnung ist daher nur mehr oder weniger genau über Hilfsgrößen (wie z.B. die Kombination von Namen und Geburtsdatum) möglich.

(5) Die Aufwertung des Transparenzregisters zu einem Vollregister – also unter Einbezug derjenigen Daten, die eigentlich schon an anderer Stelle verfügbar sein müssten – ist ein wichtiger Schritt. Allerdings gelten aktuell noch Übergangsfristen. Grundsätzlich bestehen bleibt das Problem, dass die Einträge aus Selbstauskünften der Berechtigten beruhen. Immerhin wird die Transparenz erhöht.

(6) Geldwäscheverdachtsmeldungen werden normalerweise nicht öffentlich bekannt. Indes belegen partielle Enthüllungen, dass dieser Datenschatz allenfalls sehr bruchstückhaft gehoben wird. Es besteht anscheinend ein gravierendes Vollzugsdefizit.

(7) Bedauerlich ist, dass bereits innerhalb Deutschlands die öffentlichen Register nicht die Bedingungen für „Offene Daten“ im Sinne der Open Knowledge Foundation erfüllen: *Data available as a whole; at no more than a reasonable reproduction cost; in a convenient and modifiable form*. Abfragen lassen sich in der Regel nur einzelne Datensätze und damit entstehen erhebliche Kosten. Die Tatsache, dass öffentliche Register von einem privaten (privatisierten!) Unternehmen betrieben werden, ist ein Schildbürgerstreich ersten Grades. Dies sollte dringend geändert werden, wobei realistischer Weise höchstens die rechtlichen Entwicklungen auf EU-Ebene Anlass zu Hoffnung bietet.

(8) Entsprechende öffentliche Registerdaten wären zumindest in der gesamten EU erforderlich. Das würde die Datenerhebungen für die Finanzverwaltungen aller Staaten erheblich erleichtern und zielgerichteter machen. Immerhin bieten Staaten wie Luxemburg, Niederlande, Irland, Malta und Zypern bewusst Möglichkeiten für legale Steuergestaltungen an, die aber auch in illegaler Weise missbraucht werden können. Nicht mehr Mitglied der EU ist Großbritannien. Allerdings erlaubt das dortige „Companies House“ bislang sehr weitreichende Datenzugriffe.

(9) Je weniger Offene Daten vorliegen, desto aufwändiger und potenziell fehlerträchtiger ist die Sammlung und Aufbereitung der Daten. Diese Arbeiten kann die Finanzverwaltung natürlich selbst durchführen. Jedoch wäre für bestimmte Daten der Zukauf von schon fertig aufbereiteten Datensammlungen von kommerziellen Anbietern eine Alternative. Dieser Markt ist stark im Wachsen begriffen. Sollte es nicht möglich sein, komplette Datensätze zu kaufen, so stellt *Robotic Process Automation* (RPA) einen Weg für massenhafte automatisierte Einzelabfragen dar.

(10) Für die Sammlung, Integration und Analyse der Alternativen Daten bietet es sich an, auf etablierte und erprobte internationale Standards zurückzugreifen. Nach Auffassung der Autoren eignet sich am besten eine flexible Graphen-Datenbank in der Version eines RDF-Stores. RDF gehört zu den technischen Standards der EU als Element ihrer Digitalisierungsstrategie. Das RDF-Konzept ist auch Teil der internationalen

Standards des W3C (World Wide Web Consortium) für das Internet. Im Unterschied zu (vielen) privaten Datenformaten fallen hierbei keine Lizenzgebühren an.

(11) Es lassen sich im Zeitvergleich drei eindeutige Trends erkennen: Erstens nimmt die Menge der potentiell verfügbaren Alternativen Daten permanent und exponentiell zu. Zweitens gibt es verstärkte Bemühungen staatliche Daten als Offene Daten bereit zu stellen. Und drittens greift der Gedanke von „verknüpften" Linked Open Data langsam um sich. Damit ist ein wohl unumkehrbarer Prozess in Gang gesetzt, der die Nutzbarkeit von Alternativen Daten für die Betriebsprüfung stetig ausweiten wird.

(12) Jedoch müssen diese Potenziale auch gehoben werden. Dafür braucht es den Willen, ein entsprechendes System von Menschen und Technik einzurichten. Und last not least, auch Kapazitäten und Knowhow seitens der Betriebsprüfer (und Steuerfahnder), den entdeckten Verdachtsmomenten nachzugehen.

Literatur

AADE (2020): myDATAElectronic Books AADE - Technical documentation of REST API interface for submittingand retrieving data Version 0.6 – March2020, https://www.aade.gr/sites/default/files/2020-04/myDATA%20API%20Documentation%20v0%206b_eng.pdf (8.9.2020).

Alderman, Liz / Schuetze, Christopher (2020): In a German Tech Giant's Fall, Charges of Lies, Spies and Missing Billions, The New York Times vom 26.06.2020, https://www.nytimes.com/2020/06/26/business/wirecard-collapse-markus-braun.html (01.07.2020).

Alexiev, Vladimir (2019): euBusinessGraph Semantic Data Model, Version 1.4., https://docs.google.com/document/d/1dhMOTlIOC6dOK_jksJRX0CB-GIRoiYY6fWtCnZArUhU/edit#heading=h.dskoqyysgc2l (17.9.2021).

Allemang, Dean / Hendler, Jim (2011): Semantic Web for the Working Ontologist, 2. Aufl., Morgan Kaufmann, Waltham 2011.

Amel-Zadeh, Amir / Serafeim, George (2018): Why and How Investors Use ESG Information: Evidence from a Global Survey, Financial Analysts Journal, Vol. 74(3), S. 87-103.

Arnaout, Hiba / Elbassuoni, Shady (2018): Effective searching of RDF knowledge graphs, Journal of Web Semantics 48, S.66-84.

Auer, Sören / Pellegrini, Tassilo / Sack, Harald (2014): Linked Enterprise Data: Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien. Springer-Verlag.

Balbierer, Thomas et al. (2021): Neues Steueroasen-Leak belastet Hunderte Politiker, 3.10.2021, https://projekte.sueddeutsche.de/artikel/politik/pandora-papers-geheimgeschaefte-von-politikern-enttarnt-e500259/ (8.10.2021).

Balogh, Krisztian (2018): Entity-Oriented Search, Springer Open Verlag, https://rd.springer.com/content/pdf/10.1007%2F978-3-319-93935-3.pdf (22.9.2021).

Bartz, Tim / Böcking, David / Hesse, Martin / Traufetter, Gerald (2020): Wirecard-Skandal: Wirtschaftsprüfer EY und Aufsichtsbehörde machen sich gegensei-

tig Vorwürfe, Der Spiegel vom 09.12.2020, https://www.spiegel.de/wirtschaft/unternehmen/wirecard-skandal-wirtschaftspruefer-ey-und-aufsichtsbehoerde-machen-sich-gegenseitig-vorwuerfe-a-f011387f-8e19-4c82-99f7-037b27076b3f (23.12.2020).

Beck, Wolf (2021): Reform des GwG - Transparenzregister wird zum Vollregister, Haufe News vom 27.7.2021, https://www.haufe.de/compliance/recht-politik/reform-des-gwg-transparenzregister-wird-vollregister_230132_548132.html (17.9.2021).

Becker, Markus / Diehl, Jörg / Traufetter, Gerald (2021): Scholz überging Warnungen zu Geldwäsche-Fahndern, Spiegel Online vom 10.9.2021, https://www.spiegel.de/wirtschaft/zoll-spezialeinheit-fiu-olaf-scholz-ueberging-warnungen-zu-geldwaesche-fahndern-a-4f9ef6fc-faf6-48e1-9a25-7a3350a526ae (118.9.2021).

Berliner Senat, Verordnung zur Bereitstellung von allgemein zugänglichen Datenbeständen (Open Data) durch die Behörden der Berliner Verwaltung (Open Data Verordnung - OpenDataV) vom 24.Juli 2020.

Berlinonline (2021): Das Berliner Open-Data-Handbuch, https://berlinonline.github.io/open-data-handbuch/ (13.3.2021).

Berners-Lee, Tim (2009): Linked Data, https://www.w3.org/DesignIssues/LinkedData.html (31.7.2021).

Bittner, Thomas / Dawid, Roman / Metzner, Susann (2016): Typische Problemfelder in Betriebsprüfungen. In: Dawid, Roman (Hrsg.) Verrechnungspreise. Springer Gabler, Wiesbaden 2016, https://doi.org/10.1007/978-3-658-09377-8_6 (25.9.2021).

BMF (2019): Erste Nationale Risikoanalyse. Bekämpfung von Geldwäsche und Terrorismusfinanzierung. 2018/2019, https://www.bundesfinanzministerium.de/Content/DE/Downloads/Broschueren_Bestellservice/2019-10-19-erste-nationale-risikoanalyse_2018-2019.pdf?__blob=publicationFile&v=17 (18.9.2021).

BMF, Schreiben zur Mitteilung grenzüberschreitender Steuergestaltungen vom 29.3.2021, V A 3 - S 0304/19/10006 :010.

Bundesrechnungshof (2020): Bericht nach § 99 BHO über Maßnahmen zur Verbesserung der Umsatzsteuerbetrugsbekämpfung – Chancen der Digitalisierung nutzen, https://www.bundesrechnungshof.de/de/veroeffentlichungen/produkte/sonderberichte/langfassungen-ab-2013/2020/massnahmen-zur-verbesserung-der-umsatzsteuerbetrugsbekampfung-chancen-der-digitalisierung-nutzen Bonn, 29.10.2020.

Bundeszentralamt für Steuern (2020): Aktuelle Liste der teilnehmenden Staaten. Stand: 1. Juli 2020, https://www.bzst.de/SharedDocs/Downloads/DE/CRS/crs_teilnehmende_staaten_2020.pdf?__blob=publicationFile&v=12 (8.3.2021).

Bundeszentralamt für Steuern (2021a): Common Reporting Standard – CRS, https://www.bzst.de/DE/Privatpersonen/Selbstauskuenfte/CommonReportingStandard/commonreportingstandard_node.html (8.3.2021).

Bundeszentralamt für Steuern (2021b): Kommunikationshandbuch - Automatischer Austausch von Steuergestaltungen (DAC 6), https://www.bzst.de/DE/Unternehmen/Intern_Informationsaustausch/DAC6/Handbuecher/handbuecher.html#js-toc-entry1 (25.9.2021).

Cloudflare (2021): How CAPTCHAs work | What does CAPTCHA mean?, https://www.cloudflare.com/de-de/learning/bots/how-captchas-work/ (23.10.2021).

Cohen, Jeffrey / Holder-Webb, Lori / Khalil, Samer (2017): A Further Examination of the Impact of Corporate Social Responsibility and Governance on Investment Decisions, Journal of Business Ethics, Vol. 146, S. 203-218.

COSO (Committee of Sponsoring Organizations of the Treadway Commission) (2013): Internal Control - Integrated Framework (updated), https://www.coso.org (12.3.2021).

Creative Commons (2021a): About The Licenses, https://creativecommons.org/licenses/?lang=en (31.7.2021).

Creative Commons (2021b): What is Creative Commons and what do you do?, https://creativecommons.org/faq/#what-is-creative-commons-and-what-do-you-do (31.7.2021).

Christian, Christopher / Liebscher, Marc / Wortham, Leah (2020): Wirecard, Europe's Enron? - Auditor Liability to Investors Corporate Fraud, Other Lectures & Events. 33, https://scholarship.law.edu/other_lectures/33/ (13.12.2020).

Dhaliwal, Dan S. / Li, Oliver Zhen / Tsang, Albert / Yang, Yong George (2011): Voluntary Nonfinancial Disclosure and the Cost of Equity Capital: The Initiation of Corporate Social Responsibility Reporting, The Accounting Review, Vol. 86(1), S. 59-100.

Diehl, Jörg (2021): Länder erneuern Kritik an Zoll-Spezialeinheit, https://www.spiegel.de/panorama/justiz/zoll-spezialeinheit-fiu-kampf-gegen-geldwaesche-laender-erneuern-kritik-a-48f32de2-2288-4c34-88bf-a786ef17d16d (8.10.2021).

Dietrich, Daniel (2011): Was sind offene Daten, http://www.bpb.de/gesellschaft/digitales/opendata/64055/was-sind-offene-daten (12.3.2021).

Drüen Klaus-Dieter in Tipke/Kruse, AO/FGO, 154. EL 10.2018, AO § 138a Rz. 14.

Eccles, Robert G. / Kastrapeli, Mirtha D. / Potter, Stephanie J. (2017): How to Integrate ESG into Investment Decision-Making: Results of a Global Survey of Institutional Investors, Journal of Applied Corporate Finance, Vol. 29(4), S. 125-133.

Ehrlinger, Lisa / Wolfram Wöß (2016): Towards a Definition of Knowledge Graphs, SEMANTiCS, 48(1-4), 2.

Endt, Christian / Munzinger, Hannes / Obermaier, Frederik / Prugger, Daniela / Wormer, Vanessa (2019): Der Eigentümer bleibt geheim, Süddeutsche Zeitung, 5.2.2019, https://www.sueddeutsche.de/wirtschaft/transparenzregister-firmeneigentuemer-eu-1.4317342 (13.3.2021).

Engert, Markus (2020): Was sind die FinCEN-Files?, BuzzFeedNews vom 20.9.2020, https://www.buzzfeed.com/de/marcusengert/was-sind-die-fincen-files (18.9.2021).

Engert, Markus / Drepper, Daniel (2020): Die FinCEN-Files: Wie Großbanken an Oligarchen, Drogendealern und Terroristen verdienen, BuzzFeedNews vom 20.9.2020, https://www.buzzfeed.com/de/marcusengert/fincen-files-banken-verdienen-an-kriminellen (18.9.2021).

EU, Durchführungsbeschluss (EU) 2017/1358 der Kommission vom 20. Juli 2017 zur Festlegung der technischen Spezifikationen im IKT-Bereich, auf die bei der Vergabe öffentlicher Aufträge Bezug genommen werden kann, Amtsblatt der Europäischen Union L 190 vom 21.7.2017, S. 16-19.

EU, Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen „Aufbau eines gemeinsamen europäischen Datenraums“ vom 25.4.2018, COM (2018) 232 final.

EU, Rat der Europäischen Union, Schlussfolgerungen des Rates zur Mitteilung der Kommission über eine externe Strategie für effektive Besteuerung und Empfehlung der Kommission zur Umsetzung von Maßnahmen zur Bekämpfung des Missbrauchs von Steuerabkommen, 9452/16, FISC 85, ECOFIN 502 vom 25.5.2016.

EU, Rat der Europäischen Union, Schlussfolgerungen des Rates zur überarbeiteten EU-Liste nicht kooperativer Länder und Gebiete für Steuerzwecke (2021/C 66/10), Amtsblatt der Europäischen Union C 66 vom 26.2.2021, S. 40-45.

EU, Richtlinie 2014/55/EU des Europäischen Parlaments und des Rates vom 16. April 2014 über die elektronische Rechnungsstellung bei öffentlichen Aufträgen, Amtsblatt der Europäischen Union L 133, 6.5.2014, S. 1-11.

EU, Richtlinie 2014/95/EU des Europäischen Parlaments und des Rates vom 22. Oktober 2014 zur Änderung der Richtlinie 2013/34/EU im Hinblick auf die Angabe nichtfinanzieller und die Diversität betreffender Informationen durch bestimmte große Unternehmen und Gruppen, Amtsblatt der Europäischen Union L 330 vom 15.11.2014, S. 1-9.

EU, Richtlinie (EU) 2015/849 des Europäischen Parlaments und des Rates vom 20. Mai 2015 zur Verhinderung der Nutzung des Finanzsystems zum Zwecke der Geldwäsche und der Terrorismusfinanzierung, zur Änderung der Verordnung (EU) Nr. 648/2012 des Europäischen Parlaments und des Rates und zur Aufhebung der Richtlinie 2005/60/EG des Europäischen Parlaments und des Rates und der Richtlinie 2006/70/EG der Kommission, Amtsblatt der EU L 141 vom 5.6.2015, S. 73-117.

EU, Verordnung (EU) Nr. 596/2014 des Europäischen Parlaments und des Rates vom 16. April 2014 über Marktmissbrauch (Marktmissbrauchsverordnung) und zur Aufhebung der Richtlinie 2003/6/EG des Europäischen Parlaments und des Rates und der Richtlinien 2003/124/EG, 2003/125/EG und 2004/72/EG, Amtsblatt der Europäischen Union L 173 vom 12.6.2014, S. 1-61.

EU, Verordnung (EU) Nr. 600/2014 des Europäischen Parlaments und des Rates vom 15. Mai 2014 über Märkte für Finanzinstrumente und zur Änderung der Verordnung (EU) Nr. 648/2012, Amtsblatt der Europäischen Union L 173 vom 12.6.2014, S. 84-148.

EU Kommission (EU) (2016) 198: Vorschlag für eine RICHTLINIE DES EUROPÄISCHEN PARLAMENTS UND DES RATES zur Änderung der Richtlinie 2013/34/EU im Hinblick auf die Offenlegung von Ertragsteuerinformationen durch bestimmte Unternehmen und Zweigniederlassungen.

Fensel, Dieter et al. (2020): Knowledge Graphs, Springer Nature Switzerland Verlag, Cham 2020.

Foote, Keith D. (2018): What Is a Document Database? http://www.dataversity.net/tag/document-store/ (17.10.2021).

Fuchs, Christina / Steiner, Gerhard (2016): Verrechnungspreisdokumentationsgesetz: Hinweise zum länderbezogenen Bericht SWI 2016, 388.

Gesetz zur Einführung einer Pflicht zur Mitteilung grenzüberschreitender Steuergestaltungen, Regierungsentwurf v. 9.10.2019 https://dip.bundestag.de/vorgang/.../253781 (17.10.2021).

GLEIF (2017): Common Data File Formats – Questions and Answers, https://www.gleif.org/de/about-lei/common-data-file-format# (19.9.2021).

Grotherr, Siegfried (2016): Anwendungsfragen bei der länderbezogenen Berichterstattung – Country-by-Country Reporting IStR 2016, 991-1008.

Grümmer, Julian (2021): What can we learn from knowledge graphs? - A Wirecard perspective, Vortrag auf der Knowledge Graph Conference 2021, May 3-6 2021, https://knowledgegraphconference.vhx.tv/packages/kgc-2021/videos/julian-grummer-what-can-we-learn-from-knowledge-graphs-a-wirecard-perspective (24.10.2021).

Gwerder, Zoe (2020): Zug und die Briefkastenfirmen: die Schattenseiten des Erfolgs, Luzerner Zeitung, 4.4.2020, https://www.luzernerzeitung.ch/zentralschweiz/zug/zug-und-die-briefkastenfirmen-die-schattenseiten-des-erfolgs-ld.1210007 (25.2.2021).

Haarmann, Wilhelm (1977): Abgabenordnung - Einführung und Begriffsbestimmungen, Inf. 1977.

Harle, Georg / Nüdling, Lars / Olles, Uwe (2020), Die moderne Betriebsprüfung, 4., aktualisierte und erweiterte Auflage, NWB Verlag, Herne 2020.

Henneberger, Bernd (2012): Die Abgrenzung des Konsolidierungskreises unter IFRS, Shaker Verlag, Aachen 2012.

Henselmann, Klaus / Haller, Stefanie (2018): Potentielle Risikofaktoren in der E-Bilanz-Taxonomie für die Erhöhung der Betriebsprüfungswahrscheinlichkeit, Deutsches Steuerrecht DStR 20/2018, S. 1033-1039.

Henselmann, Klaus / Hofmann, Stefan (2010): Accounting Fraud. Case Studies and Practical Implications, Erich Schmidt Verlag, Berlin 2010.

Henselmann, Klaus / Schmidt, Lutz / Sigloch, Jochen (2005): Internationale Steuerlehre. Steuerplanung bei grenzüberschreitenden Transaktionen, Gabler Verlag, Wiesbaden 2005.

Hodler, Amy / Needham, Mark (2021): Graph Data Science (GDS) for dummies, Wiley, Hoboken 2021.

Hofmann, Stefan (2008): Handbuch Anti-Fraud-Management, Erich Schmidt Verlag, Berlin 2008.

Horák, Josef /Bokšová, Jiřina /Strouhal, Jiří (2020): Electronic Invoicing Adoption within the European Union, International Advances in Economic Research, Volume 26, S. 449–450.

Huang, Feiqi / Vasarhelyi, Miklos A. (2019): Applying robotic process automation (RPA) in auditing: A framework, International Journal of Accounting Information Systems, Volume 35, 2019, 100433, DOI: 10.1016/j.accinf.2019.100433.

Kafsak, Hendrik / Schäfers, Manfred (2021): Konzerne müssen Gewinn für jedes EU-Land veröffentlichen, FAZ vom 25.2.2021, https://www.faz.net/aktuell/wirtschaft/country-by-country-reporting-konzerne-muessen-gewinn-fuer-jedes-eu-land-veroeffentlichen-17216429.html (12.3.2021).

Kahle Holger / Schulz Sebastian (2016): BEPS-1-Gesetz: Einführung einer drei-stufigen Verrechnungspreisdokumentation DStZ 2016, S. 823-824.

Kaya, Devrimi / Seebeck Andreas (2019): The dissemination of firm information via company register websites – Country level empirical evidence, Journal of Accounting and Organizational Change, 15(3), S. 382–429.

Kejriwal, Mayank (2019): What Is a Knowledge Graph?, Domain-Specific Knowledge Graph Construction, Springer International Publishing, S. 1–7.

Kendall, Elisa F. / McGuinness, Deborah L. (2019): Ontology Engineering. Morgan & Claypool, Williston 2019.

Kiryakov, Atanas (2016): Linked Leaks: A Smart Dive into Analyzing the Panama Papers, https://www.ontotext.com/blog/linked-leaks-a-smart-dive-into-analyzing-the-panama-papers/ (17.9.2021).

Kowallik, Andreas / Eismayr, Rainer / Kirsch, Andreas (2016): Globale Entwicklungen bei der Automation von Besteuerungsprozessen, Der Betrieb, Beilage 04 zu Heft Nr. 47, 25.11.2016, S. 40-46.

Krahnen, Jan Pieter / Langenbucher, Katja (2020): The Wirecard lessons: A reform proposal for the supervision of securities markets in Europe, No. 88, SAFE Policy Letter, https://www.econstor.eu/handle/10419/222230 (23.10.2021).

Kroetsch, M / Weikum, Gerhard (2016): Special issue on knowledge graphs, *Journal of Web Semantics*, Vol. (37)38, S. 53-54.

Lobbypedia (2020): Eintrag „Gerhard_Schröder", https://lobbypedia.de/wiki/Gerhard_Schröder (25.2.2021).

Lutz, Fabian / Seebeck, Andreas (2019a): Country-by-Country Reporting: Herausforderungen und Möglichkeiten einer automatisierten ersten Risikoeinschätzung, Internationales Steuerrecht (IStR), 14/2019, S. 535-543.

Lutz, Fabian / Seebeck, Andreas (2019b): US-amerikanische Bundessteuerbehörde IRS veröffentlicht aggregierte Country-by-Country Reports: Erste Berechnungen und Hinweise auf Problematiken bei der Analyse, Steuer & Wirtschaft International (SWI) 9/2019, S. 438-449.

Lutz, Fabian / Seebeck, Andreas (2020): OECD veröffentlicht aktualisierte Guidance zum Country-by-Country Reporting, Internationales Steuerrecht (IStR) 29/2020, S. 55-59.

Lutz, Fabian (2020) Eignung des Country-by-Country Reportings der OECD zur Einschätzung von ausgewählten BEPS-Risiken und Ableitung eines Ansatzes zur Verbesserung des Country-by-Country Reportings, Duncker & Humblot, Berlin 2020.

Ministry of Economics, Greece (2020): The myDATA platform for electronic pricing is ready, https://www.vatupdate.com/wp-content/uploads/2020/06/2020-06-13-mydata-en.pdf, 12. Juni 2020 (8.9.2020).

Monk, Ashby / Prins, Marcel / Rook, Dane (2019): Rethinking Alternative Date in Institutional Investmet, The Journal of Financial Data Science, Winter 2018, S. 14-31.

Moffitt, Kevin C. / Rozario, Andrea M. / Vasarhelyi, Miklos A. (2018): Robotic Process Automation for Auditing, Journal of Emerging Technologies in Accounting, Vol. 15, No. 1, Spring 2018, S. 1-10, DOI: 10.2308/jeta-10589.

Nayak, Ameya, Poriya, Anil / Poojary, Dikshay (2013). Type of NOSQL databases and its comparison with relational databases. International Journal of Applied Information Systems, 5(4), S. 16-19.

Obermayer, Bastian / Obermaier, Frederik (2016): Panama Papers. Die Geschichte einer weltweiten Enthüllung, 2. Aufl., Kiepenheuer & Witsch, Köln 2016.

OECD (2017): Country-by-Country Reporting: Handbook on Effective Tax Risk Assessment, 29.9.2017, www.oecd.org/tax/beps/country-by-country-reporting-handbook-on-effective-tax-risk-assessment.pdf (17.10.2021).

OECD (2018): Guidance on the Implementation of Country-by-Country Reporting, BEPS Action 13 (Guidance), https://www.oecd.org/tax/beps/guidance-on-country-by-country-reporting-beps-action-13.htm (13.9.2018).

OECD (2021): What is BEPS? (12.3.2021).

OECD (2010), OECD Transfer Pricing Guidelines for Multinational Enterprises and Tax Administrations 2010 (TPG), OECD Publishing, Paris 2010.

Offeneregister.de (2021): German company register data: README, https://offeneregister.de/daten/ (13.3.2021).

OFR (2020): Was Wien für Briefkastenfirmen attraktiv macht, ORF.at, 22.6.2020, https://orf.at/stories/3170200/ (25.2.2021).

Open Knowledge Foundation (2021): What is open?, https://okfn.org/opendata/ (31.7.2021).

Opencorporates (2019): German company data now open for all, https://blog.opencorporates.com/2019/02/05/german-company-data-now-open-for-all/ (13.3.2021).

Oxfam (2016c): Data in Excel File, https://oi-files-d8-prod.s3.eu-west-2.amazonaws.com/s3fs-public/file_attachments/tb-race-to-bottom-methodology-spreadsheet-121216-en.xlsx (7.3.2021).

Oxfoam (2016a): Tax Battles. The dangerous global Race to the Bottom on Corporate Tax, https://oi-files-d8-prod.s3.eu-west-2.amazonaws.com/s3fs-public/bp-race-to-bottom-corporate-tax-121216-en.pdf (7.3.2021).

Oxfoam (2016b): Technical Methodology Document. How Oxfam identified the world's worst corporate tax havens, https://oi-files-d8-prod.s3.eu-west-2.amazonaws.com/s3fs-public/file_attachments/tb-race-to-bottom-methodology-note-121216-en.pdf (7.3.2021).

Paulheim, Heiko (2017): Knowledge graph refinement: A survey of approaches and evaluation methods, Semantic web, Vol. 8(3), S. 489-508.

Peemöller, Volker H. / Kregel, Joachim (2010): Grundlagen der internen Revision, Erich Schmidt Verlag, Berlin 2010.

Peemöller, Volker H. / Krehl, Harald / Hofmann, Stefan / Lack, Jana (2020): Bilanzskandale, 3. Aufl., Erich Schmidt Verlag, Berlin 2020.

Redeker, Steffen (2020): Der DAX in Steueroasen. Studie für die Fraktion DIE LINKE. im Bundestag, https://www.fabio-de-masi.de/de/article/2757.studie-der-dax-in-steueroasen.html (3.3.2021).

Russel, Matthew / Klassen, Mikhail (2019): Mining the Social Web, 3. Auflage, O'Reilly Verlag, Sebastopol 2019.

Sauer, Otto (1988): Steuerliche Außenprüfung, Vahlen Verlag, München 1988.

Schaefer-Drinhausen, Heinrich (2021): Steuerhinterziehung bei privaten Verkäufen, https://www.raekoeln.de/steuerhinterziehung-bei-privaten-verkaufen/ (13.3.2021).

Schneider, Katharina (2020): Per Airbnb zum Steuerhinterzieher: Warum Vermieter jetzt dringend handeln müssen, Handelsblatt, 16.09.2020 (https://www.handelsblatt.com/finanzen/steuern-recht/steuern/kurzvermietung-per-airbnb-zum-steuerhinterzieher-warum-vermieter-jetzt-dringend-handeln-muessen/26192082.html (13.3.2021).

Schönwitz, Daniel (2014): Mit diesen Tricks machen Sie sich fast zum Hoeneß, Wirtschaftswoche, 10.3.2014, https://www.wiwo.de/finanzen/steuern-recht/steuerhinterziehung-mit-diesen-tricks-machen-sie-sich-fast-zum-hoeness/9583386-all.html (13.3.2021).

Seebach, Nils (2018): Bundesanzeiger 2.0 oder Stalking für Firmen, https://digitalkaufmann.de/entrepreneur-radar/bundesanzeiger-2-0-stalking-fuer-firmen/ (17.9.2021).

Seebeck, Andreas / Kaya Devrimi (2021): The Power of Words: An Empirical Analysis of the Communicative Value of Extended Auditor Reports, European Accounting Review, im Erscheinen.

Seebeck, Andreas / Lutz, Fabian (2021): Das XML-basierte Country-by-Country-Reporting - Neuerungen, Möglichkeiten und Herausforderungen, Steuer und Wirtschaft International – Tax and Business Review 5/2021, S. 256–268.

Seebeck, Andreas / Vetter, Julia (2021): Not Just a Gender Numbers Game – How Board Gender Diversity Affects Corporate Risk Disclosure, Journal of Business Ethics, im Erscheinen.

Sharma, Vatika / Dave, Meenu (2012): SQL and NoSQL Databases, International Journal of Advanced Research in Computer Science and Software Engineering 2(8), S. 20-27.

Shaxson, Nicholas (2012): Treasure Islands: Tax Havens and the Men who Stole the World, Vintage Verlag, London 2012.

Singhal, Amit (2012): Introducing the Knowledge Graph: things, not strings, 16.5.2012, https://blog.google/products/search/introducing-knowledge-graph-things-not/ (8.10.2021).

Swanson, Ana (2016): How the U.S. became one of the one of the world's biggest tax havens, The Washington Post, 5.3.2016, https://www.washingtonpost.com/news/wonk/wp/2016/04/05/how-the-u-s-became-one-of-the-worlds-biggest-tax-havens/ (8.3.2021).

Tax Justice Network (2021a): Our History, https://www.taxjustice.net/our-history/ (7.3.2021).

Tax Justice Network (2021b): Financial Secrecy Index, https://fsi.taxjustice.net/en/ (7.3.2021).

Uschold, Michael (2018): Demystiying OWL for the Enterprise, Morgan & Claypool, Williston 2018.

Véron, Nicolas (2020): The Wirecard debacle calls for a rethink of EU, not just German, financial reporting supervision, Bruegel vom 30.06.2020, https://www.bruegel.org/2020/06/the-wirecard-debacle-calls-for-a-rethink-of-eu-not-just-german-financial-reporting-supervision/ (23.12.2020).

von Brocke, Klaus / Nonnenmacher, Roland / Przybilka, Stefan (2021): Anzeigepflichten für grenzüberschreitende Steuergestaltungen, NWB Verlag, Herne 2021.

von Daniels, Justus / Wörpel, Simon (2019): Mieten unter Palmen, https://correctiv.org/aktuelles/wem-gehoert-hamburg/2019/02/05/mieten-unter-palmen/ (13.3.2021).

Wortham, Leah / Liebscher, Marc / Christian, Christopher D. (2020), Wirecard, Europe's Enron? - Auditor Liability to Investors Corporate Fraud, https://scholarship.law.edu/other_lectures/33 (24.10.2021).

Zucman, Gabriel (2015): The Hidden Wealth of Nations, The University of Chicago Press, Chicago 2015.

Section B.2

Analyzing the quality of iXBRL company accounts in the UK

Working Paper

Accepted for presentation at:

52nd British Accounting and Finance Association Annual Conference 2020, Southampton, United Kingdom

Contents – Section B.2

Analyzing the quality of iXBRL company accounts in the UK

Abstract

This study closes the existing gap in literature by investigating the quality of iXBRL reports from small and medium-sized enterprises (SME) in the UK in the years 2016–2019. I therefore identify the eight most common errors which occur in my dataset and evaluate them on the basis of existing literature. Results show that the three most common errors are the wrong structure of the tags, the incorrect reference of tags and the formatting of the tags. Overall, the results show, that the quality of the iXBRL accounts is very good and that most iXBRL filings do not include errors. This result is especially interesting. The UK is the only country where the filing of iXBRL accounts of SME companies is mandatory. In addition, this study focuses on SME companies and proves that the company size does not necessarily determine the quality of the iXBRL filing.

Keywords

eXtensible Business Reporting Language; XBRL; Information Quality; Data Quality

1 Introduction

Over time, everything evolves. Not only companies have changed, but so has the way they provide information about their business. Financial reporting has benefited, at least for the addressees of these annual reports. Increasingly efficient digitization, for example, makes it possible to automatically evaluate annual reports. Gone are the days when annual reports were only available in PDF format and preferably for reading only. Unfortunately, the addressee could not do anything with these annual reports. It was very effortful and time-consuming to evaluate information and, if necessary, transfer it to a database.

Data processing has changed significantly in some respects over the years. However, there have been changes especially in the processing side and less in the creation of the data. With the introduction of the eXtensible Business Reporting Language (XBRL), the creation of data has also evolved. XBRL is a text-based format for exchanging information in the context of electronic financial reporting. iXBRL (Inline XBRL) is the further development of XBRL instance documents. iXBRL documents can be read by machines as well as by humans. This means that software is no longer required to read the XBRL document without any problems. In addition, the reporting company has the possibility to design the documents individually.

In the UK, the new electronic reporting format iXBRL was introduced relatively early. Since April 2011, companies in the UK are required to file their company tax returns, including financial accounts and computations, in iXBRL. However, there are several inconsistencies and errors that can occur in the reporting process which also affect the quality of the report.

Data is an essential resource for companies. Prior literature shows that the use of XBRL in companies reduces cost (Blankespoor, 2019), increases efficiency (Dhole et al., 2015; Amin et al., 2018; Chen & Zhou, 2019; Du & Wu, 2019), and lowers complexity (Cong et al., 2019; Hoitash & Hoitash, 2018; Li & Nwaeze, 2018). However, literature towards the quality of iXBRL company accounts is lacking.

This paper closes the existing gap in literature by investigating the quality of filed iXBRL accounts from small and medium-sized enterprises (SME) in the years 2016–2019 in the UK. In total, I analyze 882.796.471 iXBRL tags from 2.892.841 companies. Most of the companies (81,20 %) in the dataset are micro companies.

Furthermore, most of the companies (60,09 %) have less than 100.000 GBP, but more than 10.000 GBP in terms of their total assets. Most of the companies are located in the south of the UK. Almost every second iXBRL filing is created using IRIS Accounts Production Software (built-in).

To further analyze the quality of the iXBRL filings, I identify the eight most common errors which occur in the dataset. My results show that the most common error which occurs in 9.639.583 tags is a missing field value. For example, it is likely that a company has forgotten to tag a value or has not tagged it properly. The second most common error which was found in 3.044.048 tags is an incorrect statement of time. The third most common error which appears in 1.038.675 tags is the wrong structure of the tags. This error may appear in various forms. For instance, the structure of the tags may have been interchanged.

Overall, the results show, that the quality of the iXBRL accounts is very good and that most iXBRL filings do not include errors. This result is especially interesting. The UK is the only country where the filing of iXBRL accounts of SME companies is mandatory. In addition, this study focuses on SME companies and proves that the company size does not necessarily determine the quality of the iXBRL filing.

This paper is organized as follows. The second chapter contains the theoretical background and gives an overview of the basics of iXBRL, XBRL in the UK, the existing conceptual frameworks of data and information quality, the creation of iXBRL filings and the common errors of XBRL implementation. Chapter 3 includes the research methodology by presenting the data retrieval, the data sample and the descriptive analysis. The results focus on the eight most common errors in the data sample and are presented in Chapter 4. The paper ends with a conclusion in chapter 5.

2 Literature review & theoretical background

2.1 Literature Review

In general, the established literature highlights several uses for XBRL, such as monitoring financial institutions, risk management, transparency, accountability, reducing the administrative burden of financial information disclosure (Liu, 2013; Perdana et al., 2015).

From a company's internal point of view, the use of XBRL reduces cost (Blankespoor, 2019), increases efficiency (Dhole et al., 2015; Amin et al., 2018; Chen & Zhou, 2019; Du & Wu, 2019), and lowers complexity (Cong et al., 2019; Hoitash & Hoitash, 2018; Li & Nwaeze, 2018).

Since data is an essential resource for companies, its quality is very important. For this purpose, various dimensions can be considered:

Gallagher (1974) highlighted four important aspects of data and information quality (DIQ): relevance, informativeness, usefulness, and importance. According to Ahituv's (1980) research, these can be extended with the following dimensions: accuracy, timeliness, aggregation, and formatting (Ahituv, 1980). Wang and Strong (1996) and IASB (2015) provide a very detailed review of the dimensions of DIQ. Perdana et al. (2018) compared these two frameworks and combined them into a DIQ framework for XBRL filings. This framework serves as a basis for quantitative analysis.

Data is an essential resource for companies. Prior literature shows that utilizing XBRL in companies reduces cost (Blankespoor, 2019), increases efficiency (Dhole et al., 2015; Amin et al., 2018; Chen & Zhou, 2019; Du & Wu, 2019), and lowers complexity (Cong et al., 2019; Hoitash & Hoitash, 2018; Li & Nwaeze, 2018). However, literature towards the quality of iXBRL company accounts is lacking.

Some researchers already focused on certain criteria to analyze the DIQ: For example, Dhole et al. (2015) analyzes the effects on the SEC’s XBRL mandate on financial reporting comparability. Debreceny et al. (2010) figure out that a quarter of 435 XBRL filings in the United States contained calculation errors. However, according to Bartley et al. (2011), the error rate in XBRL filings has decreased since the first XBRL implementation in the U.S. This may be explained by the experiential learning of companies (Du et al., 2013).

Hodge et al. (2004) and Arnold et al. (2010) in turn, point out that XBRL improves users' experience of browsing and searching for information and finding relevance material in financial reports.

However, literature towards the quality of SME XBRL company accounts in the UK is lacking. The goal of this paper is to evaluate the quality of iXBRL filings in the UK in the years 2016-2019 by investigating the eight most common errors.

2.2 XBRL & iXBRL: General information

XBRL is the open international standard for digital business reporting managed by XBRL International. It is an XML-based format that allows companies to exchange information quickly, accurately, and digitally. What makes XBRL so reliable is the tagging of the information, where reported items are clearly identified (An Introduction to XBRL, 2021). In addition to financial statements, XBRL can be used in other business reports, such as financial information, non-financial information, general ledger transactions, regulatory filings, annual and quarterly reports, and risk and performance reports (The Standard for Reporting, 2018; Extensible Business Reporting Language (XBRL) 2.1, 2003).

The benefits of using XBRL can be seen in the example of a department in the state of Nevada, where the controller finds that XBRL meets the goals of timely and accurate data, stronger internal controls, reduced costs, and a standardized system for seamless data exchange (Hoffmann & Rodríguez, 2013). Currently, there are XBRL use cases in more than 50 countries (An Introduction to XBRL, 2021).

Due to the need to publish financial and business information in both machine-readable and human-readable formats, the Inline XBRL specification has been extended from the XBRL 2.1 standard. The current version is Inline XBRL Specification 1.1, which has been implemented since 2013 and is a second version of the standard (Inline XBRL Part 1: Specification 1.1, 2013). While HTML is used to represent data and describe the structure of a web page, XML is used to store and exchange data. Similarly, iXBRL is developed as an alternative way of exchanging XBRL data with report representations. In other words, iXBRL is an HTML document with XBRL tagged data embedded in it.

Unlike XBRL documents, reports in iXBRL format require no special tools and can be opened in any web browser. In addition, XBRL software can be used to extract the XBRL data in the iXBRL report. A direct link between numbers and text in the report display and the values of the XBRL facts are essential features of iXBRL (iXBRL Tagging Features, 2019).

2.3 The use of XBRL in the UK

The amount of open accessible data varies from jurisdiction to jurisdiction in terms of volume, veracity, velocity and variety. The UK Companies House, for instance, publishes account data on a daily basis whereas the SEC publishes the financial statement of approx. 9.000 companies on a quarterly basis (XBRL International).

Since April 2011, companies in the UK are required to file their company tax returns, including financial accounts and computations, in iXBRL. The appropriate filing software is available through HMRC (Her Majesty's Revenue Customs) since November 2009. The software enables smaller companies with less complex tax affairs to file their returns in the correct format. Until now, the software market offers a range of conversion/tagging solutions.

Each year, around 1.9 million companies are now filing their information in iXBRL to the tax authority, HRMC. The complexity of the accounts varies from complex reports for large organizations to simple ones from small companies. As there is no prescribe layout, they vary in terms of format and presentation. Most of the reports are filed under UK GAAP standards. Only the large companies file under International Financial Reporting Standards (IFRS).

To make the introduction of XBRL easier, the HMRC initially set a reduced tagging requirement. Instead of using the full set of tags, UK companies were allowed to use a reduced set of tags. But still, HMRC encouraged to use the full tagging. Under the latest taxonomies, all kind of companies must file fully tagged data.

In 2013, Companies House started to publish free-of-charge iXBRL accounts. Especially investors, financial institutions and the public makes use of the published data. Since June 2015, the UK Companies House provides a public beta service that gives free access to over 170 million company records. The service can be accessed through a web service and an application program interface. This enables consumers and software provider's real time updates.

2.4 Existing conceptual frameworks of data and information quality

In literature, the terms “Data Quality” and “Information Quality” are often used interchangeably (Lee et al., 2002; Neely & Cook, 2011). In information systems (IS), data and information quality (DIQ) are a significant area of research (Wang & Strong, 1996; Wand & Wang, 1996; Lee et al., 2002).

Gallagher (1974) developed four important aspects of DIQ: relevance, informativeness, usefulness and importance. Further research developed the aspects and added accuracy, timeliness, relevance, aggregation and formatting (Ahituv, 1980). Wang & Strong (1996) developed a DIQ framework in IS that’s consists of 15 dimensions: access security, accessibility, accuracy, appropriate amount of data, believability, completeness, concise, consistency, ease of understanding, interpretability, objectivity, reputation, relevancy, timeliness and value added. These 15 dimensions can be classified in four categories, namely intrinsic, contextual, representational, and accessible (Wand & Strong, 1996). Intrinsic DIQ refers to the inherent quality of the information itself (i.e. accuracy, believability, objectivity, and reputation). Contextual DIQ is associated with the fitness of the data and information to support the task being undertaken (i.e., appropriate amount of data, completeness, relevancy, timeliness, and value added). Representational DIQ denotes that the information is easy to understand and process by data consumers (i.e., concise, consistency, ease of understanding, and interpretability), while accessible DIQ refers to IS’ support of data and information security (i.e., access security and accessibility) (Wang & Strong, 1996; Lee et al., 2002).

Table 1 shows the DIQ dimensions of Wang and Strong (1996), the description of the dimension and the classification.

DIQ dimension	Description of the dimension	Classification
Access Security	The extent to which the quantity or volume of available data and information is appropriate	Accessibility
Accessibility	The extent to which data and information are available or easily and quickly retrievable	Accessibility
Accuracy	The extent to which data and information are correct, reliable and certified free of error	Intrinsic
Appropriate Amount of Data	The extent to which quantity or volume of available data and information is appropriate	Contextual
Believability	The extent to which data and information are accepted or regarded as true, real and credible	Intrinsic
Completeness	The extent to which data and information are well documented, verifiable and easily attributed to a source	Contextual
Concise	The extent to which data and information are compactly represented without being overwhelming (i.e., brief in presentation, yet complete and to the point)	Representational
Consistency	The extent to which data and information are always presented in the same format and are compatible with the previous format	Representational
Ease of Understanding	The extent to which data and information are clear, without ambiguity and easily comprehended	Representational
Interpretability	The extent to which data and information are inappropriate language and units and the data definitions are clear	Representational
Objectivity	The extent to which data and information are unbiased (unprejudiced) and impartial	Intrinsic
Relevancy	The extent to which data and information are applicable and helpful for the task at hand	Contextual
Reputation	The extent to which data and information are trusted or highly regarded in terms of their source or content	Intrinsic
Timeliness	The extent to which the age of the data is appropriate for the task at hand	Contextual
Value Added	The extent to which data and information are beneficial and provide advantages for their use	Contextual

Table 1: DIQ dimension of Wang and Strong (1996)

In accounting, the concept of DIQ is similar as in the field of IS (Neely & Cook, 2011). Here, data and information are regarded as useful if they fulfil the criteria of useful financial reporting. Therefore, the International Accounting Standard Board (IASB) (2015) developed a concept of DIQ consisting of 11 dimensions. The IASB framework is analogous to the framework of Wang and Strong (1996) presented above.

Table 2 shows the dimensions, definitions and categories of the IASB DIQ framework.

Dimensions	Definitions	Categories
Relevance	The extent to which data and information can affect decision making	Core qualitative characteristics of financial reporting
Materiality	The extent to which data and information affect decision making when the data or information are omitted or misstated	Component of relevance
Faithful representations	The extent to which data and information represent an economic phenomenon	Core qualitative characteristics of financial reporting
Complete	The extent to which data and information are well depicted with all necessary depictions and explanations	Component of faithful representations
Neutrality	The extent to which data and information are free from bias and free from favorable and unfavorable manipulation	Component of faithful representations
Prudence	The extent to which data and information are provided with cautions	Component of faithful representations
Free from errors	The extent to which data and information are free from errors or omissions of description and explanation	Component of faithful representations
Comparable	The extent to which data and information enable decision makers to understand and to identify similarities and differences	Additional qualitative characteristics
Verifiable	The extent to which data and information can be proved as correct when presenting the economic phenomena	Additional qualitative characteristics
Timely	The extent to which data and information are available for decision makers in time and appropriate to influence decision making	Additional qualitative characteristics
Understandable	The extent to which data and information are readily comprehended	Additional qualitative characteristics

Table 2: Data and information quality of the IASB (2015)

Perdana et al. (2019) compared the DIQ framework for IS and the DIQ framework in accounting. The result is a DIQ framework for XBRL filings. This framework serves as a basis for quantitative analysis. Table 3 by Perdana et al. (2019) compares the different DIQ frameworks.

DIQ Framework in IS (Wang and Strong, 1996)	DIQ Framework in Accounting (IASB, 2015)	Preliminary DIQ Framework for XBRL
Access Security		Access Security
Accessibility		Accessibility
Accuracy	Free from errors	Accuracy
Appropriate Amount of Data		Appropriate Amount of Data
Believability		Believability
Completeness	Complete	Completeness
Concise		Concise
Consistency		Consistency
Ease of Understanding	Understandability	Ease of Understanding
Interpretability		Interpretability
Objectivity		Objectivity
Relevancy	Relevance	Relevancy
Reputation		Reputation
Timeliness	Timeliness	Timeliness
Value Added		Value Added
	Comparability	Comparability
	Faithful representation	Faithful representation
	Materiality	Materiality
	Neutrality	Neutrality
	Prudence	Prudence
	Verifiable	Verifiable

Table 3: Preliminary DIQ framework for XBRL by Perdana et al. (2019)

2.5 Creation of iXBRL filings

iXBRL company accounts can be generated with two major approaches: built-in and bolt-on. Getting to know the theories behind these two approaches makes it easier to understand the difficulties and possible error areas.

2.5.1 Built-in approach

Basically, there are two approaches for the in-house implementation of iXBRL: the built-in and the bolt-on approach. Since on the one hand, there is no indication for the establishment of a so-called "form-based solution" and on the other hand, outsourcing

to an external provider represents a temporary solution, both possibilities will not be considered in the following.

When using an integrated approach (Built-in Approach), XBRL tagging is integrated into the company's creation process (Berger & Lieck, 2018). This requires a reporting software package that enables the mapping of financial statement data to XBRL taxonomy items and the generation of iXBRL documents. Figure 1 shows an example of the creation process when using the Built-in Approach. The required data is imported via interfaces to various data sources such as the company's ERP system or Microsoft Office applications.

The data contained in the taxonomy is assigned to the elements of the taxonomy by means of mapping and in the case of company-specific features, new elements (extensions) are added to the taxonomy. If the software does not support fully automated mapping, the user must perform some of this manually. In most software solutions, mapping is necessary once and can be reused in subsequent years (Garbellotto, 2009).

The financial report can then be generated in various formats such as PDF, Word or iXBRL. The validation of the finished iXBRL documents is largely carried out using automated controls.

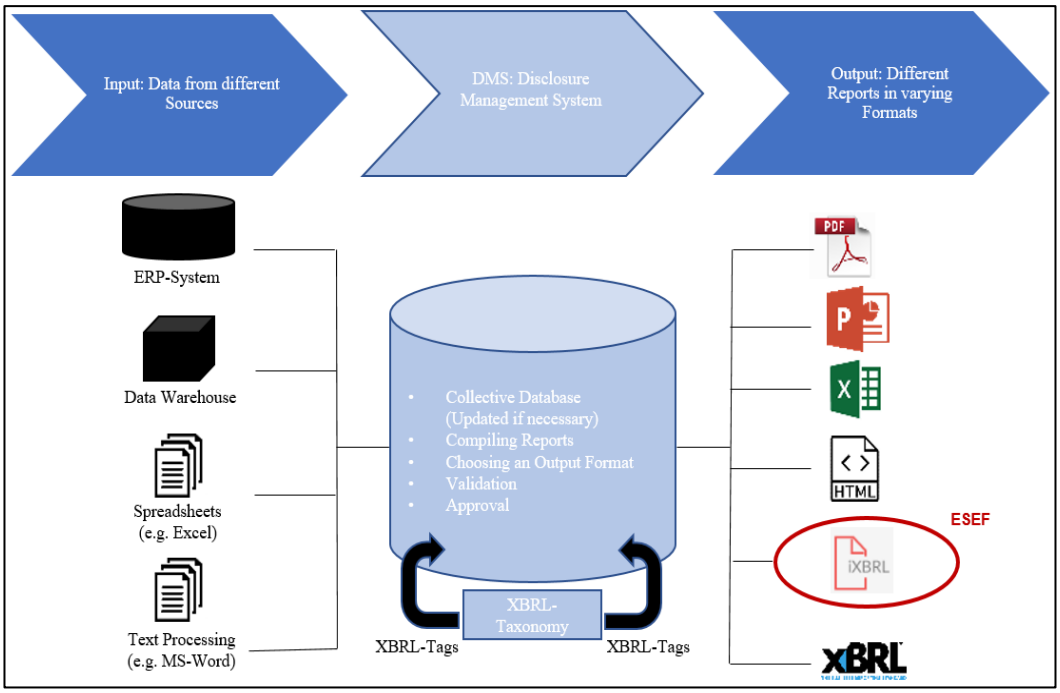


Figure 1: Built-in approach by Henselmann et al. (2019)

For companies that do not use integrated software, the implementation of the built-in approach is both time and cost intensive, as the previous creation process needs to be redesigned. If the implementation is not carried out by a provider, detailed XBRL expertise is required. These costs in the beginning are offset by lower running costs in subsequent years (ESMA, 2016).

Thus, the high degree of automation enables the reporting process to run smoothly, thus saving time and effort. This effect is particularly high for groups and companies with complex structures. Furthermore, adjustments in subsequent years are only necessary in the event of legal changes such as accounting standards. A further advantage is the centrally stored database, which ensures consistency of data throughout the report and increases transparency. One can easily make updates and if one works with the General Ledger (GL) taxonomy, the tagged closing items can be traced down to the underlying transactions. The reduction of manual interfaces and implementation of automated controls also reduces the susceptibility to errors, which in turn increases the quality of the financial reports. Adopting different data sources and IT systems also opens new possibilities, since e.g., the financial statement data of subsidiaries abroad

that use other IT systems can be imported without any problems. Furthermore, an integrated approach allows the data to be used for various company-specific reporting purposes.

In addition to the annual financial report in iXBRL format, risk reports or reports to the management and supervisory board can be generated for internal reporting purposes. With an integrated approach, companies also benefit from the advantages of structured data such as the new analysis options resulting from the machine readability and evaluability.

2.5.2 Bolt-on approach

If the bolt-on approach is used, the financial report is first prepared in the format used by the company. In a second step, the XBRL tags are then assigned to the finished report components (Garbellotto, 2009). Figure 2 visualizes the bolt-on approach.

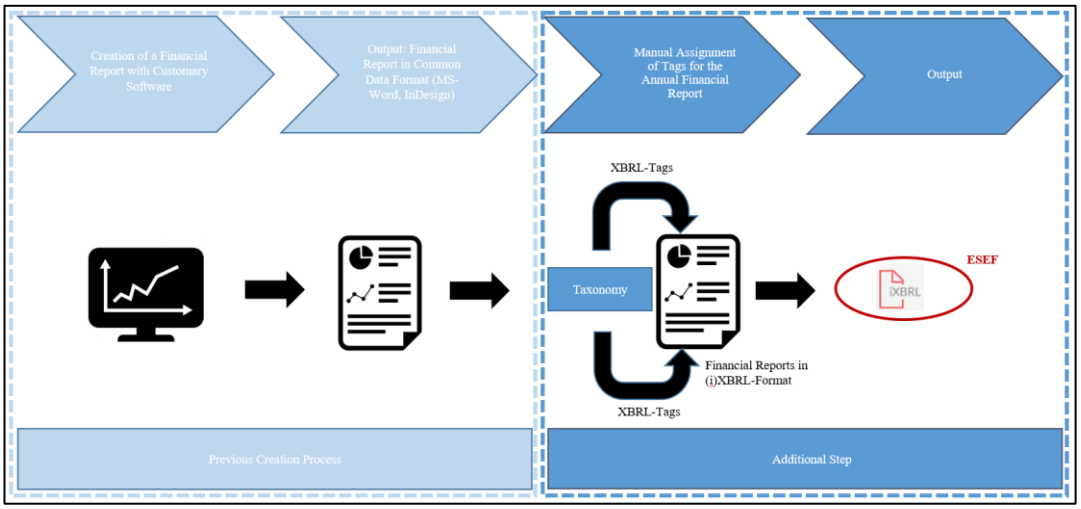


Figure 2: Bolt-on approach by Henselmann et al. (2019)

Bolt-on software solutions can be Microsoft Office add-ins, stand-alone products into which the various report components are imported, or web-based technology (Software-as-a-service, SaaS). In the latter case, the software provider makes its software available online and takes care of configuration, maintenance and updates. Another advantage over licensed software is more flexible pricing.

For mapping, the numbers or text fields of the financial report are marked and assigned to the corresponding elements of the taxonomy. Extensions may have to be created for company-specific features.

The taxonomy is either stored in the software or is stored by the user. The assignment is made manually by using drag and drop. Advanced software solutions make mapping proposals based on an algorithm that evaluates the structure of the report components and uses text analysis methods to assign the appropriate taxonomy position. There is also self-learning software that remembers the taxonomy positions selected in the past. After mapping or tagging, the iXBRL report can be generated.

Bolt-on software solutions add an extra step to the standard creation process, making implementation less time and costly compared to an integrated approach. While this allows for timely implementation, the additional step of tagging must be repeated each year (ESMA, 2009).

To ensure the correctness of the mapping or tagging, manual controls must be introduced. In addition, auditing by an auditor is more costly than with an integrated approach, where a system-check and a check of the extensions is sufficient. Compared to the built-in approach, this procedure involves higher costs in subsequent years. Furthermore, the use of different systems carries risks regarding the consistency of the data and is more costly for companies in the long term than centralized data storage. Depending on the degree of automation, the user needs to acquire detailed knowledge about iXBRL functionality and mapping or tagging. In contrast to build-in software solutions, the data cannot be used for different reporting purposes such as internal reporting (Garbellotto, 2009).

The bolt-on approach is therefore particularly suitable for companies that need the software exclusively to meet the new legal requirements and companies that cannot implement an integrated approach due to time constraints. Companies that cannot afford an integrated approach, for example because the previous creation process was very complex, also benefit from the bolt-on approach.

2.6 Common errors of XBRL implementation

There are several inconsistencies and errors that can arise in the reporting process. Depending on how the report was created, there are different error sources. Accordingly, the following section first discusses the different types of report creation. Then the process of creating the report is examined in more detail. The process can be divided into different steps.

Table 4 shows the four steps of developing a XBRL document by Bartley et al. (2010).

Process Steps	**Common Errors**
Mapping	- Selecting inappropriate elements from the U.S. GAAP Taxonomy - Creating unnecessary new elements
Extensions	- Presenting elements in the wrong location in the rendered financial statements - Failing to establish mathematical relationships among elements
Tagging	- Assigning incorrect signs to the value of the elements - Making data-entry errors
Creating and validating	- Failure to validate adequately XBRL documents both manually and with validation software

Table 4: Steps of developing a XBRL document and common errors

Mapping is the process of identifying and reconciling each accounting concept and the corresponding amount in a company's financial statements with the corresponding financial statement element in the XBRL U.S. GAAP taxonomy. Enterprises have made numerous mapping errors, some of the most serious, as they distort the meaning of data downloaded into analytical software. Users may not be able to identify financial statement concepts that have been matched against incorrect XBRL elements without detailed comparisons of XBRL elements with the original financial statements and notes.

Various forms of aberrations continue to be observed under the SEC mandate. In the observations of SEC staff, this matter is referred to as "element selection". The submitters have used standard elements in the taxonomy that are either too wide or too narrow to fit the actual balance sheet concept exactly. For example, the submitter selects the "Inventory, finished goods" element when a narrower standard element, "Energy-related inventory, petroleum," is available to capture the meaning of the underlying financial reporting concept more accurately.

A more common and generally more serious misstatement is the creation of unnecessary new elements that "extend" the U.S. GAAP taxonomy.

This occurs when a preparer fails to find the correct element in the taxonomy and creates a new, unique element for a financial reporting concept.

I have also observed the unnecessary creation of new, duplicate elements when a single concept appears in more than one place in financial reporting. For example, if preferred dividends appear in the income statement, cash flow and equity, the correct allocation for all three locations is to the standard element in the U.S. GAAP taxonomy for "dividends, preferred stock, and cash". XBRL provides the coding to display a single element in multiple locations with different names where appropriate.

The flexibility to create unique elements is one of the strengths of XBRL and can seem harmless. However, analytical software recognizes only the standard elements in the US GAAP taxonomy, so each unique element requires some degree of manual interpretation by interactive users. In addition, extension and labeling errors are much more likely to occur with unique elements because companies must enter all the information that would otherwise be automatically provided for XBRL for standard elements, such as data type, period type, debit or credit balance, definition, and labeling.

The key is to create new elements through the enhancement process only when an annual financial statement concept is not included in the standard taxonomy. Organizations can request that XBRL US add new elements to the official taxonomy through the public commenting process that is followed for all new versions of the taxonomy.

The extension process creates new XBRL elements in the taxonomy that contain the essential information needed to produce a company's unique financial statements. This computer code is complex, and until now, most companies have relied on third parties to perform the extension process. The XBRL code allows a company to control or establish unique presentation labels for elements. The location of elements within the financial statements, including multiple locations. Mathematical relationships that allow the sum of amounts within a related group to be validated. For example, the amounts of elements in the group for current assets should add up to the amount entered for the current assets subtotal. Other information required to establish the formal structure of the financial statements of a particular enterprise. For example, display descriptions such as "current assets" that are not linked to an element.

Extension errors often cause serious errors in financial statements and can distort the interpretation of XBRL data entered in analytical software.

Tagging is the process of entering both numeric and textual data for financial report elements, including dollar amounts, signs, periods, and units of measure. Tagging errors are less common than mapping and extension errors, but they are serious because the incorrect data distorts both the rendered financial reports and the data downloaded into analytical software.

Sign errors usually occur for reasons other than a simple keystroke error. The correct marking of signs is complex due to the distinction between "debit" and "credit" element values and the fact that a value can be either positive or negative depending on its location and description in the financial statements. XBRL principles require the assignment of positive signs to all elements whose values correspond to their natural debit or credit balance, but some elements lack a natural balance, such as changes in working capital accounts, which are presented in a cash flow statement. In addition, signs that are determined when entering amounts in the identification process may require manipulation in the enhancement process to achieve a correct display in the financial statements prepared.

The final steps in preparing interactive XBRL data for submission to the SEC are the creation and final validation of the XBRL instance documents, which contain all tagged financial reporting elements and associated presentation information. The final preparation of the documents is a straightforward technical process. Failure to perform proper software validation and manual validation leaves many of the errors I have seen in VFP filings undetected.

Validation software automatically checks and identifies most, but not all, violations of XBRL standards. It verifies mathematical accuracy once the extension process has established the mathematical relationships between financial reporting elements and their totals; however, as noted above, XBRL cannot currently report all financial relationships in reports. Report preparers should perform software validations using both their third-party software and validation software available on the SEC's website.

The internal assurance process is iterative and should be performed continuously during the preparation of the instance documents.

Errors tend to accumulate and trigger additional errors, so waiting until the instance documents are complete to begin validation makes the validation and correction process more difficult.

Furthermore, Bartley et al. (2011) identified six general categories of errors. This categorization helps us to classify the errors in the further analysis. Table 5 presents the error types and a description of the error by Bartley et al. (2011).

Error Type	**Description of Error**
Missing Elements	Elements (concept) appear in Form 10-K, but not in the XBRL instance documents.
Incorrect Amounts	Correct elements are in the XBRL documents, but the amount or date is incorrect (excludes sign flips).
Sign Flips	Elements coded with the incorrect sign or calculation weight. (Sign errors only in the display of an element's value classified as incorrect display errors.)
Duplicate Elements	Financial statement concepts coded more than once so that they appear more than once in the instance document.
Incorrect Elements	Financial statement concepts coded with an incorrect element in the U.S. GAAP taxonomy or unique company-specific elements created unnecessarily.
Incorrect Display	Elements are otherwise correctly coded are displayed with an erroneous label, in the wrong location, or not at all in the rendered XBRL financial statements.

Table 5: Classification of errors in XBRL instance document by Bartley et al. (2011)

Notably, the categorization of errors, especially the description of errors, does not always fit in my analysis. That is why I stick with the error types but adjusted the description of errors. The result of the adjustment can be seen in Table 6.

Error Type	Description of Error
Missing Elements	Elements from the XBRL instance documents are missing.
Incorrect Amounts	Correct elements are in the XBRL documents, but the amount or date is incorrect (excludes sign flips).
Sign Flips	Elements coded with the incorrect sign or calculation weight. (Sign errors only in the display of an element's value classified as incorrect display errors.)
Duplicate Elements	Financial statement concepts coded more than once so that they appear more than once in the instance document.
Incorrect Elements	Financial statement concepts coded with an incorrect element in the taxonomy or unique company-specific elements created unnecessarily.
Incorrect Display	Elements are otherwise correctly coded are displayed with an erroneous label, in the wrong location, or not at all in the rendered XBRL financial statements.

Table 6: Adjusted classification of errors in XBRL instance document

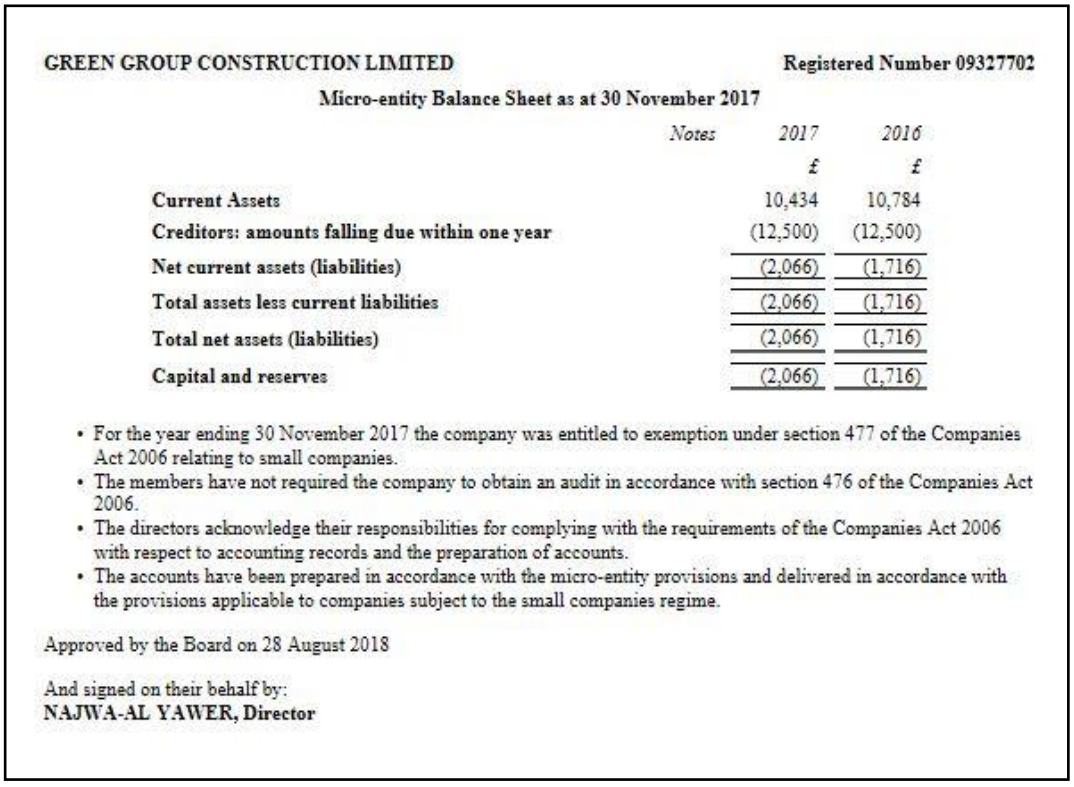
3 Research methodology

3.1 Data retrieval

All company accounts are retrieved through the website[1] of the UK Companies House. The filings are available on a daily basis as well as on a monthly basis. The individual data files are either in the iXBRL format (.html file extension) or XBRL format (.xml file extension). The available accounts are available for approx. 75 % of the 2,2 million accounts the UK Companies House expects to be filed each year. The accounts are filed in either the XML format or HTML format and can be bulk downloaded.

The type and scope of the reports differ considerably in some cases. The reports of some companies consist of two pages, the reports of other companies of 6 pages. Basically, the components of the report and the arrangement of the tags are predetermined by the UK taxonomy. Figure 3 gives an overview of an exemplary iXBRL report.

[1] http://download.companieshouse.gov.uk/en_accountsdata.html



GREEN GROUP CONSTRUCTION LIMITED **Registered Number 09327702**

Micro-entity Balance Sheet as at 30 November 2017

	Notes	*2017*	*2016*
		£	£
Current Assets		10,434	10,784
Creditors: amounts falling due within one year		(12,500)	(12,500)
Net current assets (liabilities)		(2,066)	(1,716)
Total assets less current liabilities		(2,066)	(1,716)
Total net assets (liabilities)		(2,066)	(1,716)
Capital and reserves		(2,066)	(1,716)

- For the year ending 30 November 2017 the company was entitled to exemption under section 477 of the Companies Act 2006 relating to small companies.
- The members have not required the company to obtain an audit in accordance with section 476 of the Companies Act 2006.
- The directors acknowledge their responsibilities for complying with the requirements of the Companies Act 2006 with respect to accounting records and the preparation of accounts.
- The accounts have been prepared in accordance with the micro-entity provisions and delivered in accordance with the provisions applicable to companies subject to the small companies regime.

Approved by the Board on 28 August 2018

And signed on their behalf by:
NAJWA-AL YAWER, Director

Figure 3: Micro-entity account of Green Group Construction Limited

For further analysis of the reports, it is helpful to understand the technical structure of the reports and the tags. As mentioned earlier, the structure and scope are specified by the UK taxonomy. Accordingly, these should be implemented in a fundamentally uniform manner. Only the scope can differ from company to company.

The beginning of each iXBRL account provides the used namespaces and URI's in the document. They show the specifications that determine how the document is structured. Figure 4 provides an example of the different URI's and namespaces used in an exemplary iXBRL document.

Figure 4: Example of URI's used in an iXBRL document

The namespaces are relevant for further analysis. Especially because some of the namespaces do not provide relevant information and can be neglected. For example, the namespaces with just one character (d, c, b, e, f, g). Certain namespaces, like the *ix* namespace, which is the inline XBRL part, contain most of the relevant information. Table 7 summarizes the used namespaces and the referred URI's of the company accounts.

Namespace	URI
link	http://www.xbrl.org/2003/linkbase
xbrldi	http://xbrl.org/2006/xbrldi
ixt	http://www.xbrl.org/inlineXBRL/transformation/2010-04-20
iso4217	http://www.xbrl.org/2003/iso42172
xbrli	http://www.xbrl.org/2003/instance
d	http://xbrl.frc.org.uk/cd/2014-09-01/business
xlink	http://www.w3.org/1999/xlink
c	http://xbrl.frc.org.uk/general/2014-09-01/common
b	http://xbrl.frc.org.uk/FRS-102/2014-09-01
e	http://xbrl.frc.org.uk/fr/2014-09-01/core
f	http://xbrl.frc.org.uk/reports/2014-09-01/direp
g	http://xbrl.frc.org.uk/reports/2014-09-01/aurep
ix	http://www.xbrl.org/2008/inlineXBRL
	http://www.w3.org/1999/xhtml

Table 7: Namespaces and URI used in an iXBRL document

Furthermore, for the analysis of the reports it is also necessary to understand the structure of the individual tags. Adherence to the structure of the tags is important, as this is the only way to ensure that the data can be analyzed by machine. Figure 5 shows an example of an iXBRL tag.

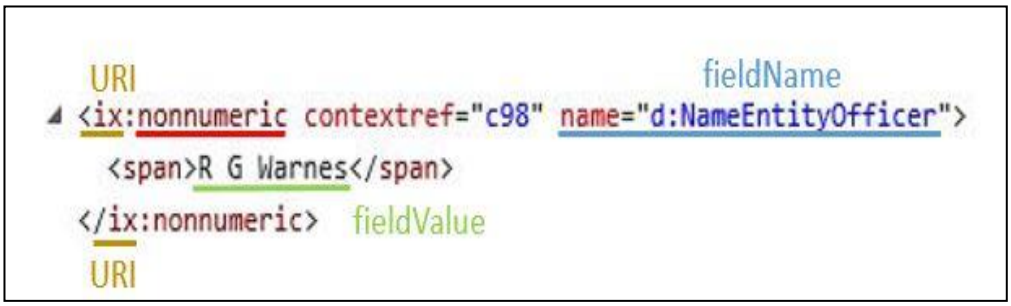


Figure 5: Exemplary *ix* tag

The *ix* in the beginning and in the end shows that it belongs to the inline XBRL namespace. The *name* section of the tag indicates the content of the tag. The following part defines whether the contained data is *nonNumeric* or *nonFraction.* Next follows the *contextref* part of the tag. In most cases, it defines the period of the tag.

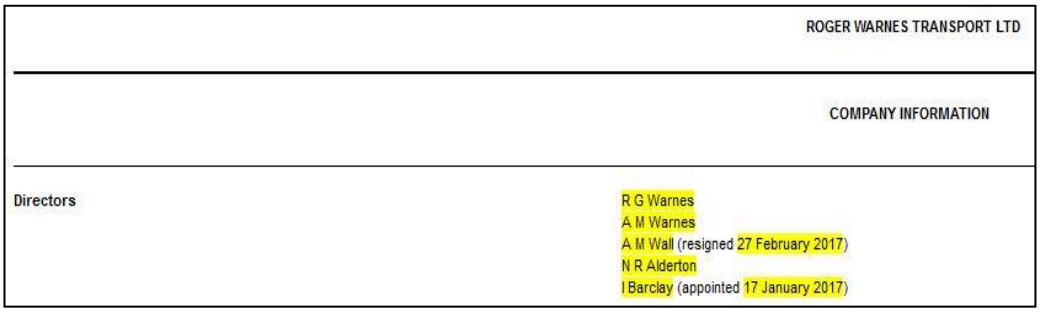
[2] ISO 4217 is an international standard for the representation of currencies (https://www.iso.org/iso-4217-currency-codes.html) (accessed 02/22/2020).

The period is either current period or the previous period. The implementation of the period differs between companies but the meaning is the same. Mostly companies state directly the year or specify e.g. “c1” for the current year and “c2” for the previous year.

In the exemplary tag in Figure 6 the name refers to the fieldname of the tag, *NameEntityOfficer*. This means the tag shows the name of the entity officer.

The tag marks a general inline element that usually gets a meaning through a class or id attribute. The real content of the tag comes after the span tag.

This part of the tag presents the *fieldValue* and expresses the value of the *fieldName*. The *fieldName* and *fieldValue* together form the basis for further analysis. R G Warnes is the content of the tag, the name of the entity officer. Figure 6 shows the tag with the name of the name of the entity officer in the human-readable part of the company account.


ROGER WARNES TRANSPORT LTD

COMPANY INFORMATION

Directors	R G Warnes A M Warnes A M Wall (resigned 27 February 2017) N R Alderton I Barclay (appointed 17 January 2017)

Figure 6: Output of exemplary tag

3.2 Data sample

I retrieve company accounts from January 2016 to December 2019. During this period 9.722.820 accounts were available. In total, the 9.722.820 company accounts belong to 2.892.841 different companies. This means that I observe in average three (3,36) filings per company. This allows a comparison of the company filings over the years.

For further analysis, all the tags from the filings are loaded into a database. In total, I have 2.139.020.400 tags. Some of the tags provided are not necessary for my further analysis. That is why I focus on the “ix” tags (instance document). Table 8 gives an overview of filings, companies and (relevant) tags.

Filings	9.722.820
Companies	2.892.841
Output (tags)	2.139.020.400
Relevant tags (“ix” = instance document)	882.796.471

Table 8: Overview of filings, companies and (relevant) tags

Table 9 in the Appendix provides an overview of the summary statistics of the monthly accounts available since January 2016. The distribution of published accounts is very similar in the different years. There is only a difference in December. In December, the accounts filed are nearly doubled to the other months.

3.3 Descriptive analysis

Even before analysing the company reports in terms of quality, it is important to understand what type of company it is. For this reason, I categorize the companies in the following regarding various criteria. These include the number of employees, the total assets, the geographical location and the production software used.

Number of employees

In order to get an overview of the size of the companies, they are classified in four different groups in terms of the number of employees.

The provided tag in the company accounts to analyze the number of employees is AverageNumberEmployeeDuringPeriod. By filtering information from the company accounts it is possible to gather data on amounts of employees for 617.079 companies. Duplicates are removed. The companies are divided into four different groups in terms of employee size. The different groups are divided into the same groups as the House of Commons uses in their Business Statistics report (Rhodes, 2018). Group one is for micro sized companies with up to 9 employees. The second group is for small sized companies with 10 to 49 employees. Group three is for medium sized companies with 50 to 249 employees and the last group contains large sized companies with 250-499 employees. Another group is added for companies with more than 500 employees. This group might potentially contain companies with tagging mistakes.

Table 10 shows the distribution of companies per class, the percentage and the cumulated percentage.

Classes	Frequency	Percentage	Cumulated %
Micro (0-9 employees)	501.062	81,20 %	81,20 %
Small (10-49 employees)	101.801	16,50 %	97,70 %
Medium (50-249 employees)	13.351	2,16 %	99,86 %
Large (249-500 employees)	508	0,08 %	99,94 %
Greater	357	0,06 %	100,00 %
Total	617.079	100,00 %	

Table 10: Distribution of companies in terms of employees

Out of the 617.079 companies, the majority (501.062) belong to group one with one to 9 employees. This amounts to around 81,20 % of all companies. Group one also contains 167.364 companies with just one employee. This means that more than one quarter (27,12 %) are one person companies. The second largest group in terms of number of employees is the second group, companies with 10-49 employees, with 16,50 % (101.801 companies). Group one and two together account for 97,70 % of all companies. The numbers are in line with the Business Statistics report by the House of Commons (Rhodes, 2018).

According to the report, the majority (96,00 %) belong to the group of micro sized companies. The second largest group is the small sized company group (4,00 %). This means that less than 3 % of the companies in the dataset have 50 or more employees. Figure 7 presents the distribution of companies in terms of employees.

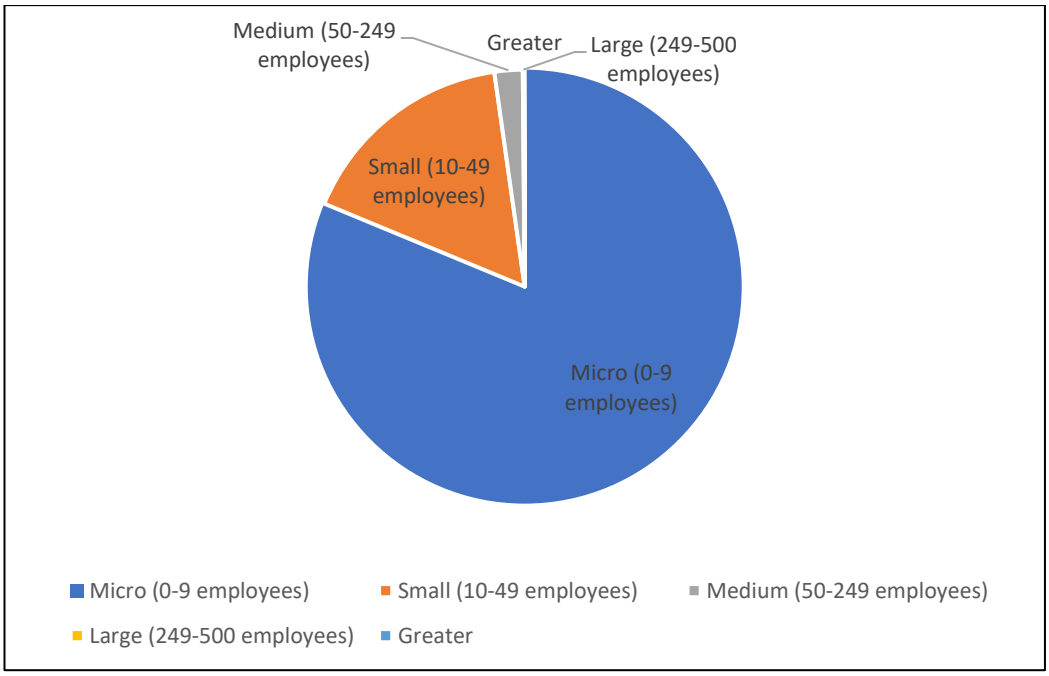


Figure 7: Distribution of companies in terms of employees

Total assets

Using financial data, the companies are classified in ten different groups in terms of their total assets. The company accounts do not provide a tag for the total assets. Therefor, the total assets are calculated by the sum of the provided tags *CurrentAssets* and *FixedAssets*. Both tags are available in the company accounts.

Table 11 shows the numbers of companies per class and their percentage. The amount of available data is limited to companies who have available data for fixed assets and current assets. Both values are needed to calculate the total assets. Any kind of duplicates are removed from the data set.

Classes (in GBP)	Group	2015	2016	2017	2018	2019	Total	Percentage
0 - 10.000	1	3.739	9.731	8.210	9.031	9.307	40.018	25,45%
10.001 - 20.000	2	1.812	4.091	3.345	3.627	3.955	16.830	10,71%
20.001 - 30.000	3	1.181	2.589	2.308	2.408	2.499	10.985	6,99%
30.001 - 50.000	4	1.678	3.590	2.924	3.265	3.502	14.959	9,52%
50.001 - 75.000	5	1.383	2.805	2.349	2.498	2.635	11.670	7,42%
75.001 - 100.000	6	1.072	1.978	1.594	1.802	2.018	8.464	5,38%
100.001 - 500.000	7	4.438	7.764	6.615	7.012	7.807	33.636	21,40%
500.001 - 1.000.000	8	1.407	2.195	1.916	1.999	2.214	9.731	6,19%
1.000.001 -2.000.000	9	758	1.285	1.174	1.221	1.266	5.704	3,63%
> 2.000.000	10	577	1.123	1.148	1.166	1.202	5.216	3,32%
		18.045	37.151	31.583	34.029	36.405	157.213	100,00%

Table 11: Total assets

The distribution of companies by total assets shows that more than one quarter of all companies (25,45 %) belong to group one, with total assets of less than 10.000 GBP. The second largest group is group 7 (21,40 %) with total assets of 100.000 GBP and 499.999 GBP. Group one and seven together amount 46,85 % of all companies. Groups 1-4 together also reflect over half (52,66 %) of the companies. This number is in line with the Business Statistics report by the House of Commons (Rhodes 2018). Their analysis shows that 99,90 % of the businesses in the UK belong to the group of SMEs, with a number of employees of 0 to 250. Figure 8 shows the distribution of the different company groups according to the classification in terms of total assets.

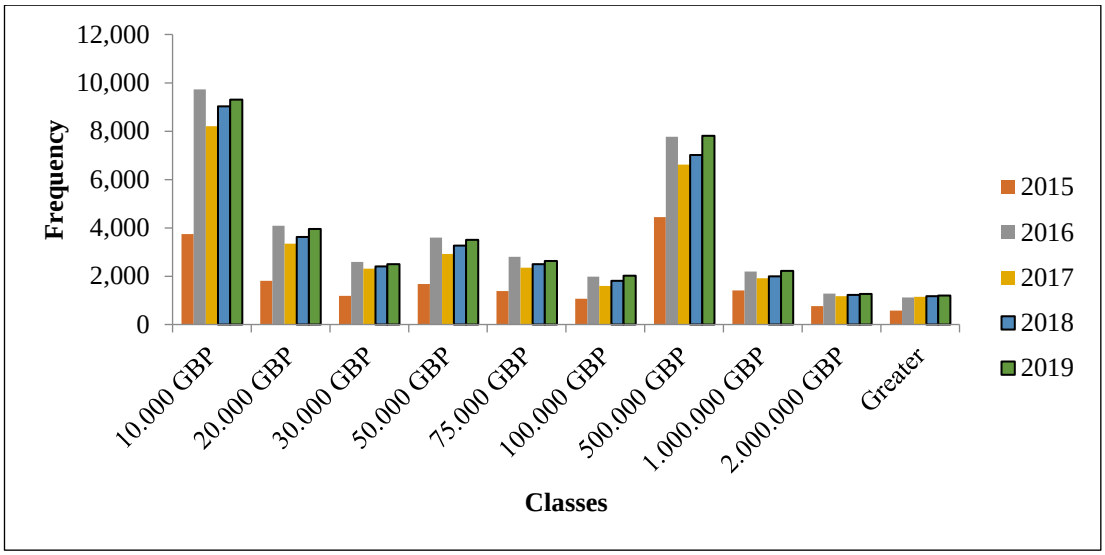


Figure 8: Distribution of companies in terms of total assets

Company location

Companies can also be classified by their geographic location. For this purpose, it is necessary that the companies provide information about their address. This does not apply to all companies. However, 373 different regions could be identified. Some company accounts provide information about the region. The relevant tag is *CountyRegion*. For all other companies, the postcode (*PostalCodeZip*) is used to assign the companies to the different regions. The geographical position of the companies is visualized in a map (Figure 9). Tableau enables a clear visualisation of the locations of the companies.

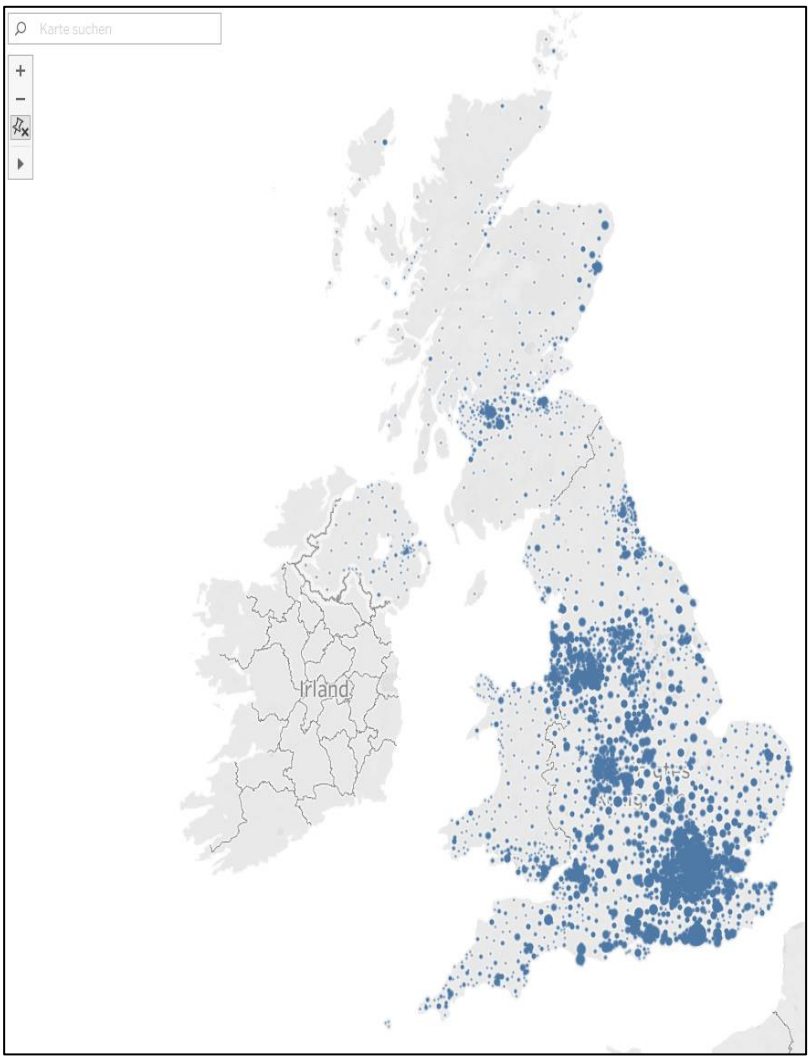


Figure 9: Geographical position of companies across the United Kingdom

The figure clearly shows that most companies are located in the south of the UK. Here, especially in London and the surrounding area. The thicker the dot, the more companies are to be found in a region. In addition, many companies can be found in the regions of Birmingham, Liverpool and Manchester.

However, it can also be seen that the identified companies are scattered throughout the country. Accordingly, it can be assumed that the data base is not limited to urban areas.

Production software used

It is also possible to identify hidden tags. This means that individual tags are included in the machine-readable part, but not in the human-readable part. An example of this would be the software used. Starting with the year 2017, companies also tag the procution software they use to create the iXBRL account.

The relevant tag to determine the production software used is *NameProductionSoftware*. By analysing the tag, it is possible to determine which companies have opted for the Built-In and which for the Bolt-On approach. To identify which software providers offer which approach, we analyzed the websites of the software providers.

The following table 12 gives an overview of the software provider used to produce the iXBRL company accounts.

Provider	**Type**	**User**	**Percentage**
Ajaccts Software	Bolt-on	90	0
Relate AccountsProduction	Bolt-on	4.050	0,10
Pinacle 6.0	Bolt-on	4.750	0,12
Capium	Bolt-on	27.820	0,71
BTCSoftware AP Solution	Built in	54.670	1,39
PTP Accounts Production	Bolt-on	65.360	1,66
Companies House	Bolt-on	91.490	2,32
Caseware UK	Built in	94.430	2,40
Taxfiler	Built in	178.880	4,54
Acorah Software Products - Accounts Production	Built in	289.980	7,36
Digita Accounts Production Advanced	Bolt-on	326.800	8,29
CCH Software	Built in	471.640	11,97
VT Final Accounts	Bolt-on	531.470	13,49
IRIS Accounts Production	Built in	1.798.500	45,65
Total		3.939.930	100,00

Table 12: Software used to create filing

The table shows the different software vendors identified and whether they are a built-in or bolt-on solution. A total of 14 different software providers were identified. There are 8 providers of bolt-on solutions and 6 with a built-in solution. For about 4 million companies it is possible to identify the software used. The remaining companies did not publish any information on which software was used to create their report. However, it should be noted that IRIS Accounts Production Software (built-in) has the largest market share in our database with 45,65 %.

4 Results

4.1 Overview

The database is very large, with well over 9 million company reports. Therefore, a reliable analysis of quality is also complicated. For this reason, I will provide a general overview of the quality of corporate reports in the first step. Afterwards, I will show the most common errors. In addition, these errors will be classified using the criteria from the literature review.

It is difficult to make a reliable statement regarding the quality of corporate reports for two reasons. Firstly, it is the sheer quantity of corporate reports.

Secondly, the quality of corporate reports is not easy to classify. In addition, it must be emphasized that some of the criteria from the Perdana et al. (2018) framework cannot be evaluated quantitatively at all, or only with difficulty. For this reason, it is important that a general perception of the reports is also analyzed. The main goal of the analysis are the questions if and how the iXBRL reports can be processed by machine. Machine processing can be limited by other errors like human evaluation. For example, spelling errors can be detected by humans and assigned appropriately. However, an algorithm will probably not recognize this error as such and for this reason will not assign the tag correctly.

Notably, the overall perception of the company accounts is very positive. Especially the processability of the company accounts is good. It is possible to download and process a large of database in a short time. The company accounts are human- and machine-readable and no problems in terms of visualization occurred. This is no wonder, as companies need to publish their company accounts for tax reasons. This might increase the motivation to hand in an error free report. Furthermore, the companies considered show a learning curve regarding the submission of their iXBRL reports. Over the years, the number of errors declined.

When comparing with the DIQ framework for XBRL by Perdana et al. (2018), it is already possible to say that the company accounts have passed the criteria of:

1) Access Security: company accounts are available through the Companies House website and easy to bulk download.
2) Accessibility: company accounts are available to bulk download.

3) Appropriate Amount of Data: number of companies and tags per company seem to be correct.
4) Concise: majority of companies used the correct structure of report and tags.
5) Consistency: majority of companies used the correct structure of report and tags.
6) Timeliness: just a small number of reports is filed for another (previous) year.

A few criteria are also difficult to evaluate. For example, due to the large database, it is difficult to make a statement as to whether the data are verifiable.

The completeness of the data was also not fully verified. For example, the analysis does not check whether the companies published all mandatory tags.

This is assumed due to the compulsory publication for tax reasons. It is also not further evaluated whether and how many extensions are used.

In summary, it can be said that the quality of the company reports is very good regarding the question of whether the reports can be evaluated automatically. Although there are a few errors that occur regularly, the entirety of the reports can be evaluated very well.

In the following, the most common errors are presented and classified with respect to the criteria from the literature section. Here, it is particularly important to understand that the errors especially complicate a machine processing of the reports. In addition, I have also classified the errors according to the stages of creation. For some of these errors, it is also possible to explain how the errors occurred. However, this is not possible for all errors. Table 13 gives an overview of the most common errors[3] and their classification.

[3] Note that neither the content nor the completeness of the tags itselfs is analyzed.

Error	Error type	DIQ XBRL Framework	In process steps	Quantity
Null-Value	Missing elements	Completeness	Mapping	9.639.583
Incorrect statement of time	Incorrect Amounts	Accuracy, Compatibility, Ease of Understanding	Tagging	3.044.048
Wrong structure of the tags	Incorrect Amounts	Accuracy	Tagging	1.038.675
Reference	Incorrect Amounts	Accuracy, Ease of Understanding	Tagging	1.248
Formatting	Incorrect Amounts	Accuracy, Compatibility, Ease of Understanding	Tagging	1.100
"?"	Incorrect Display	Comparability, Verifiable	Creating and validating	799
„ "	Incorrect Display	Comparability, Verifiable	Creating and validating	640
Error in formula	Incorrect Display	Verifiable	Extensions	496

Table 13: Overview and classification of most common errors

The total number of errors found is 13.726.589. As a result, the average number of errors per filing is 1,41 (13.726.589 errors/9.722.820 filings). This result is in line with other comparable studies that show an error rate of 0,94 (Hui et al., 2013).

4.2 Analysis of the eight most common errors

4.2.1 Null-Value

The most common and most significant error is a missing value. This can happen because of a variety of reasons. Sometimes companies simply forget to tag a value, sometimes the value is not present and sometimes the tagging was not performed properly. The result is all the same and leads to a missing value. A missing field value occurs in 9.639.583 tags of the database. However, missing values can also exist in other parts of the tag. The problem is that it is not possible to flawlessly determine whether the value has been mislabelled or is simply not there. Companies could also exploit this type of error to ultimately not publish information that they do not want to publish. Table 14 shows the frequency of missing values.

Missing value	Frequency
fieldValue	6.313.772
xmlUri	947.758
fieldname	865.168
context	760.881
Extensions	752.004

Table 14: Frequency of missing values

4.2.2 Incorrect statement of time

Each tag contains a reference to the period of the tag. The filings provide information about the current period and the previous period. Each of the tags contain information in the *context* section, whether the *fieldvalue* is about the current or previous period. In some cases, either the year of the context or the fieldValue seems to be outdated. The following Table 15 shows the occurrence of outdated years.

Year	Context	fieldValue
2010	400	102
2011	458	230
2012	662	6.048
2013	1.742	63.796
2014	130.416	2.840.194

Table 15: Occurrence of outdated years

In addition to that, some tags even mix different time stamps. In 674 tags, the year in the *context* section does not match with the year in the *fieldValue* section. Furthermore, 12 tags show a period *startDate* of 2014 and a *fieldValue* of 2008. This mismatch of the years is not explainable.

Errors concerning an incorrect or non-existent time specification lead to the fact that the information cannot be included in the database, since it is not clear to which time period it belongs. This is of course a pity and reduces the usable database. For this reason, it is important to always ensure that the time information is correct. Incorrect statement of time appeared in 3.044.048 tags of the database.

4.2.3 Wrong structure of the tags

Standardisation is an important goal of the iXBRL standard. Companies should always adhere as closely as possible to the standard. As soon as companies do not adhere to the standard, machine evaluation becomes more difficult or, in the worst case, impossible. Unfortunately, using the wrong structure in the tags is a very common error. This error can occur in various forms. For example, the structure of the tags may have been interchanged. Another error is that the values do not match the components. For example, the reference to the point is in the part that clarifies whether it is a non-numeric value or a non-fraction. However, both errors mean that the content cannot be analyzed cleanly or worst case not at all. Errors regarding the structure probably did not occur when using built-in software. These create the tags automatically. Accordingly, this error is more likely due to bolt-on software. The wrong structure occurs in 1.038.675 tags.

4.2.4 Reference

Sometimes, the tags contain the value "#REF!" which means that the reference was not properly tagged. This error occurs in two different ways. First, some tags just contain the "#REF!". This means that the company wanted to refer to another tag and it did not work properly. Second, sometimes the "#REF!" is also part of the value. An example would be the value for the tag *TangibleFixedAssetsPolicy,* which says "Depreciation has been provided at the following rates in order to write off the assets over their estimated useful lives. Equipment 15 % reducing balance #REF! #REF!". When a tag contains the "#REF!", it is most often useless because the value, it wanted to refer to, is not present. This means that this error is critical and better to avoid. Wrong references are used in 1.248 tags in the dataset.

4.2.5 Formatting

There are a variety of errors, which can be traced back to formatting errors. One example is the representation of tables in the filing. Normally, the structure of the filing should be clarified in the stylesheet section of the filing. Some companies have not implemented this properly.

An obvious error is the wrong use of bullet points. The paragraph which should be a list should be defined with the bullet points in the stylesheet.

Sometimes companies just add the "·" for a list. This is not a serious error but it shows that the creator did not use the stylesheet properly. This error occurs mainly at companies who created their statement by using a bolt-on solution. The error occurrs in 1.100 tags in the database.

4.2.6 "?"

In some tags companies have the "?"-sign. The "?"-sign is used instead of a space. An example can be seen in the tag *DescriptionShareType* with the value "4.2%?Preference". The "?"-sing is just visible in the source code, but not in the human readable filing. In the human readable filing the sign is not visible and is correctly replaced by a space. This means that the use of the "?"-sign in the tag is a not serious error, but it complicates the automatic processing. The error occurs in 799 tags in the database.

4.2.7 „ "

„ " is the programming statement for a non-breaking space. This means that the creator of the filing wanted to add a non-breaking space in the section but did not realize that it should not be used in the tag. It is not a common mistake, but it clearly shows that the filing was created by using a bolt-on approach. Most often, the error occurs in tags with more than just one word or number in it. An example is: The director acknowledges her responsibilities for complying with the requirements of the Act with respect to accounting records and the preparation of financial statements. . The error occurred in 640 tags in the dataset.

4.2.8 Error in formula

Some filings also have problems using a formula. The final result looks different or makes no sense. Sometimes it is just an error mentioned in the tag. Sometimes, the final tag has a value but also states that there is an error. The value of the tag *NameEntityOfficer* in one example leads to the value "Error in formula ->Mrs<- K Allweis". This is definitely not properly tagged. Even pointing out an error can lead to the tag not being able to be used any further. In most of the cases where a formula is not used correctly, the tag is useless. Most of the time this error occurs because there is no value at all. In some cases, also because it states "error" and one cannot be certain whether the following information is true or not. The cause of this error can have different reasons.

Either the formula was not used correctly, or not all values for the calculation of the formula are available. The error occurred in 496 tags in the database. The error did not occur in many filings because SME are usually not required to provide tags with formulas.

5 Conclusion

This paper closes an important research gap by investigating the quality of SME iXBRL filings in the UK. The quality of the iXBRL filings is especially important as they represent an essential resource for companies. In addition, false information can mislead investors in their decision-making process.

Prior literature focuses on the several uses of XBRL such as monitoring financial institutions, risk management, transparency, accountability and reducing the administrative burden of financial information disclosure (Liu, 2013; Perdana et al., 2015). The use of XBRL is associated with various benefits as it reduces cost (Blankespoor, 2019), increases efficiency (Dhole et al., 2015; Amin et al., 2018; Chen & Zhou, 2019; Du & Wu, 2019), and lowers complexity (Cong et al., 2019; Hoitash & Hoitash, 2018; Li & Nwaeze, 2018). However, the quality of iXBRL reports has not been investigated by prior literature.

The study is based on the filings of 2.892.841 UK based SME companies and covers the years 2016–2019. In total, I analyze 882.796.471 relevant tags. In order to better analyze the quality of the underlying iXBRL filings, it is important to understand which types of companies are included in the dataset.

The total number of errors found is 13.726.589. As a result, the average number of errors per filing is 1,41 (13.726.589 errors/9.722.820 filings). This result is in line with other comparable studies that show an error rate of 0,94 (Hui et al., 2013).

Interestingly, 81,20 % of the companies are micro companies. Most of the companies (60,09 %) have less than 100.000 GBP, but more than 10.000 GBP in terms of their total assets. Most of the companies are located in the south of the UK. Almost every second iXBRL filing is created using IRIS Accounts Production Software (built-in).

In order to evaluate the quality of the filings, I focus on the eight most common errors and their prevalence in the dataset. My results show that the most common error which

occurs in 9.639.583 tags is a missing field value. For example, it is likely that a company has forgotten to tag a value or has not tagged it properly. The second most common error found in 3,044,048 tags is an incorrect time specification.

The third most common error which appears in 1.038.675 tags is the wrong structure of the tags. This error can occur in various forms. For example, the structure of the tags may have been interchanged.

Overall, the results show, that the quality of the iXBRL accounts is very good and that most iXBRL filings do not include errors. This result is of particular interest.

The UK is the only country where the filing of iXBRL accounts of SME companies is mandatory. In addition, this study focuses on SME companies and proves that the company size does not necessarily determine the quality of the iXBRL filing.

This is the first study which investigates the quality SME UK based companies of iXBRL filings. The methodology and results provide a framework for future research. In Europe, the filing of iXBRL accounts is mandatory for listed companies since 2020. Therefore, it would be interesting to investigate the quality of the iXBRL accounts in Europe and to make a comparison between the different countries.

References

An Introduction to XBRL, (2021). XBRL, https://www.xbrl.org/the-standard/what/an-introduction-to-xbrl/.

Ahituv, N. (1980). A systematic approach toward assessing the value of an information system. *MIS Quaterly*, 4(4), 61-75.

Amin, K., Eshleman, J. D., & Feng, C. (2018). The effect of the SEC's XBRL Mandate on audit report lags. *Accounting Horizons*, 32(1), 1–27.

Arnold, V., Bedard, J.C., Phillips, J.R., & Sutton, S.G. (2010). The impact of tagging qualitative financial information on investor decision making: implications for XBRL, *International Journal of Accounting Information Systems*, 13(1), 2-20.

Bartley, J., Chen, Y.Y., & Taylor, E.Z. (2011). A comparison of XBRL filings to corporate 10-Ks – evidence from the voluntary filing program. *Accounting Horizons*, 25(2), 227-245.

Berger, J., & Lieck, H. (2018). Corporate Reporting mit iXBRL – Europäisches einheitliches elektronisches Berichtsformat ab 2020 –. *KoR*. 116.

Blankespoor, E. (2019). The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate. *Journal of Accounting Research*, 57(4), 919-967.

Chen, G., & Zhou, J. (2019). XBRL adoption and systematic information acquisition via EDGAR. *Journal of Information Systems*, 33(2), 23-43.

Cong, Y., Omar, A., & Sun, H. L. (2019). Does IT outsourcing affect the accuracy and speed of financial disclosures? Evidence from preparer-side XBRL filing decisions. *Journal of Information Systems*, 33(2), 45-61.

Debreceny, R., Farewell, S., Piechocki, M., Felden, C., & Gräning, A. (2010). Does it add up? Early evidence on the data quality of XBRL filings to the SEC. *Journal of Accounting and Public Policy*, 29(3), 296-306.

Dhole, S., Lobo, G.J., Mishra, S., & Pal, A.M. (2015). Effects on the SEC´s XBRL mandate on financial reporting comparability. *International journal of Accounting Information Systems*, 19, 29-44.

Du, H., & Wu, K. (2018). XBRL mandate and timeliness of financial reporting: do XBRL filings take longer. *Journal of Emerging Technologies in Accounting*, 15(1), 57-75.

Du, H., Vasarhelyi, M.A., & Zheng, X. (2013). XBRL mandate: thousands of filing errors and so what?", *Journal of Information Systems*, 27(1), 61-78.

ESMA. (2015). Consultation Paper on the Regulatory Technical Standards on the European Single Electronic Format (ESEF). 78ff.

ESMA. (2016). Feedback Statement on the Consultation Paper on the Regulatory Technical Standard on the European Single Electronic Format (ESEF).16.

Extensible Business Reporting Language (XBRL) 2.1, (2003, December 31). XBRL, https://www.xbrl.org/Specification/XBRL-2.1/REC-2003-12-31/XBRL-2.1-REC-2003-12-31+corrected-errata-2013-02-20.html.

Gallagher, C.A. (1974). Perceptions of the value of a management information system", *Academy of Management Journal*, 17(1), 46-55.

Garbellotto, G. (2009, May). XBRL Implementation Strategies: The Bolt-on Approach. *Strategic Finance*. 56.

Garbellotto, G. (2009, August). XBRL Implementation Strategies: The Built-in Approach. *Strategic Finance*. 56-57.

Henselmann, K., Vetter, J., & Mielich, S. (2019). Offene Anwendungsfragen des European Single Electronic Format (ESEF). *Wpg*. 2019.719.

Hodge, F.D., Kennedy, J.J., & Maines, L.A. (2004). Does search-facilitating technology improve the transparency of financial reporting, *The Accounting Review*, 79(3), 687-703.

Hoffman, C., & Rodríguez, M.M. (2013). Digitizing Financial Reports – Issues and Insights: A Viewpoint. IJDAR, 13.

Hoitash, R., & Hoitash, U. (2018). Measuring accounting reporting complexity with XBRL. *The Accounting Review*, 93(1), 259-287.

Hui Du, Miklos A. Vasarhelyi, Xiaochuan Zheng. (2013). XBRL Mandate: Thousands of Filing Errors and So What?. *Journal of Information Systems*. 27 (1): 61–78. https://doi.org/10.2308/isys-50399.

Inline XBRL Part 1: Specification 1.1, (2013, November 18). XBRL, https://www.xbrl.org/specification/inlinexbrl-part1/rec-2013-11-18/inlinexbrl-part1-rec-2013-11-18.html.

IASB. (2015). Conceptual framework for financial reporting". IFRS Foundation.

iXBRL Tagging Features, (2019, October 3). XBRL, https://www.xbrl.org/guidance/ixbrl-tagging-features/.

Lee, Y.W., Strong, D.M., Kahn, B.K., & Wang, R.Y. (2002). AIMQ: a methodology for information quality assessment, *Information and Management*, 40, 133-146.

Li, S., & Nwaeze, E. T. (2018). Impact of extensions in XBRL disclosure on analysts' forecast behaviour, *Accounting Horizons*, 32(2), 57-79.

Liu, C. (2013). XBRL: a new global paradigm for business financial reporting, *Journal of Global Information Management*, 21(3), 60-80.

Neely, M.P., & Cook, J.S. (2011). Fifteen years of data and information quality literature: developing a research agenda for accounting. *Journal of Information Systems*, 25(1), 79-108.

Perdana, A., Robb, A., & Rohde, F. (2019). Textual and contextual analysis of professionals' discourses on XBRL data and information quality. *International Journal of Accounting & Information Management.*

Perdana, A., Robb, A., & Rohde, F. (2018). Does visualization matter? She role of interactive data visualization to make sense of information, Australasian Journal of Information Systems, Vol. 22.

Perdana, A., Robb, A. & Rohde, F. (2015), "An integrative review and synthesis of XBRL research in academic journals", Journal of Information Systems, 29(1), 115-153.

Rhodes, C. (2018). UK Business Statistics. https://researchbriefings.files.parliament.uk/documents/SN06152/SN06152.pdf, accessed: 02/03/2022.

The Standard for Reporting, (2018). XBRL, https://www.xbrl.org/the-standard/what/the-standard-for-reporting/.

Wand, Y., & Wang, R.Y. (1996). Data quality dimension in ontological foundations. *Communications of the ACM*, 39(11), 86-95.

Wang, R.Y., & Strong, D.M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5-33.

Appendix
Table 9: Summary statistics

Year	Number of Filings												Filings (Number)	Filings (%)
	January	**February**	**March**	**April**	**May**	**June**	**July**	**August**	**September**	**October**	**November**	**December**		
2016	132.065	139.823	158.329	146.299	149.991	166.376	157.987	154.519	214.593	164.207	186.363	318.576	2.089.128	21,49
2017	180.345	151.407	182.437	149.293	171.017	188.495	176.660	171.392	231.163	188.674	202.906	287.198	2.280.987	23,46
2018	207.449	166.285	195.484	177.282	186.746	203.760	199.646	190.325	252.435	217.001	225.323	371.869	2.593.605	26,68
2019	221.334	177.584	212.415	187.482	200.297	210.169	219.113	194.581	273.220	226.691	235.549	400.665	2.759.100	28,38
Total	741.193	635.099	748.665	660.356	708.051	768.800	753.406	710.817	971.411	796.573	850.141	1.378.308	**9.722.820**	100
Total (%)	7,62	6,53	7,70	6,79	7,28	7,91	7,75	7,31	9,99	8,19	8,74	14,18	100,00	

Section B.3

A Knowledge Graph from UK Financial Statements

(with Pornchana Kveeyan & Daniel Schraudner)

Working Paper

Presented at:

53rd British Accounting and Finance Association Annual Conference 2021, Birmingham, United Kingdom

Techné: Empirical Research Methodology: Conducting Interdisciplinary Research Conference, Best Presentation Award, online

Accepted for presentation at:

17th International Conference on Knowledge Management 2022, Potsdam, Germany

Contents – Section B.3

A Knowledge Graph from UK Financial Statements

Abstract

In this paper, we take advantage of linked data technologies for financial data integration. We present our approach to transforming financial reports and information of companies and officers into a knowledge graph. Alternative solutions to dealing with taxonomies from different years and entity linking are proposed. The graph also connects to the existing knowledge bases, namely DBpedia, Open Corporates, and Wikidata. This way, the graph can continuously generate knowledge and grow over time. Subtle business and industry insights can be discovered through this graph development. The goals of this paper are graph quality, data analysis, and data visualization. These goals are set to encourage business people to better understand how the technologies can help to improve their works.The financial data in iXBRL (inline eXtensible Business Reporting Language) format and the information used in this paper are mainly extracted from Companies House in the United Kingdom. An RDF Mapping Language tool is used to perform the transformation.

Keywords:

XBRL; Knowledge Graph; Linked Open Data; RDF; SPARQL

1 Introduction

Financial reports play a vital role in the business world. Not only do they represent company's filings to meet regulatory requirements, but they may also benefit other parties in a variety of areas. Due to the fact that financial reports contain company information such as assets, liabilities, revenue, profit or loss balance, we can use this information to analyze the company's performance and benchmark against other companies in the same or across sector(s). Depending on your needs, the report's potential applications can be limitless.

However, collecting information for a single analysis can be a time-consuming and exhausting task. The task can range from scanning the report, inputting and rechecking the information, and performing an analysis. In some countries, public bodies and companies publish financial information/reports in a digital format such as PDF on their websites. To identify the benefit to consumers and publishers of data on websites, 5-Star Linked Data described by Tim Berners Lee can be used.[1] In this case, the financial report in PDF would achieve only one star for the sake of its availability on the website. From the consumer perspective, we can seek innovations that may help us to eliminate certain activities/tasks, for instance, a tool that can detect text/number in the report. Errors and rechecking tasks, however, still remain.

Nowadays, digital transformation has dominated nearly all sectors including accounting and finance, particularly financial reports. While some countries are still in an early stage of digitizing financial reporting papers, others have adopted an international open standard for digital business reporting called eXtensible Business Reporting Language (XBRL), which was developed in 1998. This XBRL format – a machine-readable format, allows us to create and exchange reporting information rapidly, accurately, and digitally. It is now used in many jurisdictions and organizations worldwide. XBRL consumers include regulators, governments, data providers, analysts, investors, and accountants (An Introduction to XBRL, 2021). Currently, the XBRL Project Directory reports 184 XBRL implementers including financial regulators, capital markets, business registrar, and tax authorities (XBRL Project Directory, 2021).

[1] https://www.w3.org/DesignIssues/LinkedData.html.

While XBRL can only be read by machine, *iXBRL or inline XBRL* has been further developed to enable both machine and human readability. Using the HTML standard (i.e., a standard markup language for contents to display on the web), we can open any iXBRL-formatted reports (e.g., a financial report, an annual report, or an internal management report) in any web browser. This provides greater benefit to both the creators of the report and the consumers (iXBRL, 2021).

Since 2011, the iXBRL format for financial reports has been implemented in the UK. The submission of annual accounts and corporate tax returns in iXBRL format is required for UK companies to the tax revenue – *Her Majesty's Revenue and Customs (HMRC)*. In addition, voluntary filing to the UK business register – *Companies House (CH)* has been encouraged. CH does not only collect but also publishes the filings to the public without any cost. CH also offers other information such as company profile and company officer information via the Application Programming Interface (API)[2] in JSON format. According to the UK government, "*the service represented a commitment to open access and transparency which can benefit everyone – whether it's entrepreneurs, taxpayers, businesses or the public sector. Better data efficiency encourages innovation, delivers better public services and stimulates growth through new revenue streams."* (Company Reporting in the UK – an XBRL Success Story, 2015.).

To the present, HMRC and CH have been responsible for the creation of over six million XBRL documents, according to XBRL UK. These are the largest XBRL filing programs in the world (XBLR in the UK, 2021).

Research motivation

As pointed out in the introduction of this paper, using the star scheme, the report in PDF only achieves one star. Both consumers and publishers can access the data, however, the data is in the form of a human- not machine-readable format. While data published as machine-readable structured data (e.g., excel) gets two stars, publishing such data in a non-proprietary format (e.g., CSV) gets three stars. Financial reports in XBRL or iXBRL format would achieve three stars for their machine-readable data and non-proprietary format where consumers can manipulate data based on their needs.

[2] https://developer.company-information.service.gov.uk/.

Four stars can be achieved by using open standards from the World Wide Web Consortium[3] (W3C) such as *Resource Description Framework[4] (RDF)* to identify things. To achieve the highest score – five stars, the reports also need to link the data to others to provide context. Therefore, the consumers can discover more data while consuming the data from the reports (Hausenblas, 2012).

For a decade, there have been growing appeals for the *XBRL* and *semantic web technologies.* In particular, a number of literatures have developed on a transformation of XBRL financial report into a graph representation – RDF. For example, several authors (García & Gil, 2010; Carretié et al., 2012; Kämpgen et al., 2014; Lee et al., 2014; O'Riain et al., 2012; Mora-Rodriguez, 2017; Li & Zhai, 2016) proposed solutions to transform XBRL reports into an RDF format. Likewise, an effort was attempted by Asimadi et al. (2017) with iXBRL financial report. In addition, by integrating the financial report with other information, a foundation for a global data ecosystem of financial and business information could be developed (O'Riain et al., 2012; Asimadi et al., 2017). In doing so, it can result in an increased level of corporate transparency (Mora-Rodriguez et al., 2017). By using the RDF as a standard data representation model recommended by W3C, it can support data merging and solve the integration and interoperability of data (RDF, 2014; SPARQL Query Language for RDF, 2008).

In summary, to date only a limited number of studies related to the transformation of iXBRL into RDF have been identified. Also, there is still no final solution/standardization on the transformation model (Kämpgen et al., 2014; Mora-Rodriguez et al., 2017). Together with the explanation in the introduction, this paper proposes a solution to transform and integrate iXBRL filings and other information by using semantic web technologies. The term 'knowledge graph' or 'graph' is used here to refer to our data integration in RDF format. Table 1 shows an example of how our knowledge graph can be beneficial for a consumer. By using the knowledge graph, time and manual efforts can be reduced for an analysis task. Research data in this paper is drawn from three main sources: iXBRL reports, CH API, and external graphs.

[3] https://www.w3.org/

[4] https://www.w3.org/RDF/

Information/Sources	iXBRL Reports	CH API	External graphs e.g., DBpedia and Wikidata	Knowledge graph
Financial data	X			X
Company profiles		X		X
Officer profiles		X		X
Industry information		X		X
Other useful information e.g. company website, social media, images, and product details			X	X

Table 1: List of information and sources in the graph

Another benefit to consumers is to eliminate data ambiguity in which it can negatively affect the decision-making process. Figure 1 illustrates this point clearly. For instance, information in the financial reports becomes clear through the search on the website i.e., the full name of Mr. J D Millet is Joel Duncan Millet. Consequently, information regarding nationality, occupation, etc. can also be revealed. By integrating the data from these two sources, the knowledge graph can lead to significant time saving and provide insights in case that the existence of the website is unknown.

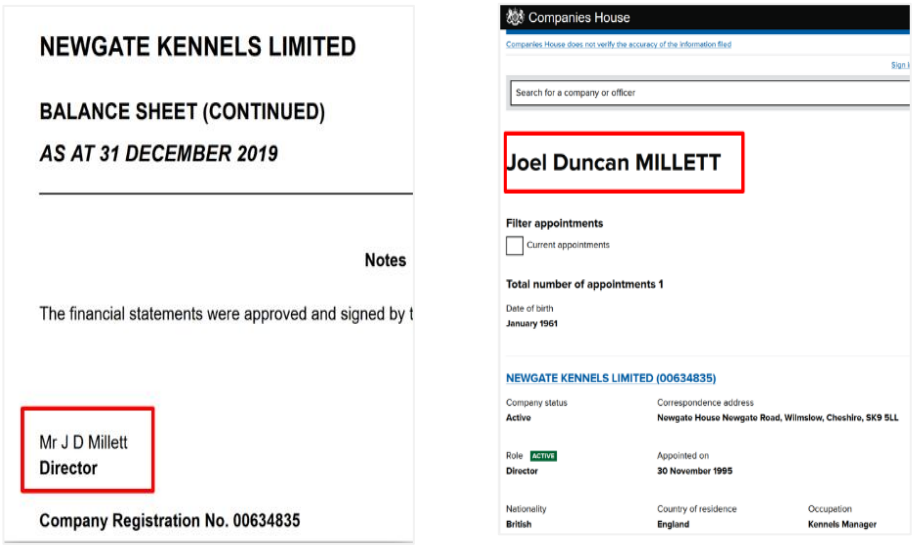


Figure 1: Example of data ambiguity between iXBRL report and CH website

Moreover, the graph also links to external graphs, namely *DBpedia*, *OpenCorporate*, and *Wikidata*.

- DBpedia[5] is a semantic version of Wikipedia. The English version of the DBpedia knowledge base currently describes 4.58 million things[6].
- OpenCorporate[7] is one of the largest open company databases in the world with almost 200 million companies in all jurisdictions worldwide. The site also provides company data in RDF format.
- Wikidata[8] is an open knowledge base that is freely accessible and edited by anyone in the world. Wikidata currently contains around 93.5 million items (Wikidata, 2021).

These well-known knowledge bases are rich in information, resources, and communities. By connecting with other knowledge bases, our graph can continuously generate knowledge and grow over time.

Research objective and contribution

Based on the abovementioned motivation, there are three goals that we aim to achieve in this paper, namely *graph quality*, *data analysis*, and *data visualization*. Based on these goals, it is hoped that this project will provide an opportunity for business people to advance an understanding of how semantic web technologies can help to improve their works.

Goal 1: Graph quality

To encourage graph usage, the users should feel confident that the graph is able to provide additional details from this data integration correctly. Many researchers have utilized *precision, recall, and F-1 scores* to measure the quality of the entity linking. The confidence can be derived from the metrics provided.

[5] https://dbpedia.org/

[6] https://wiki.dbpedia.org/about/facts-figures

[7] https://opencorporates.com/

[8] https://wikidata.org/

Goal 2: Data analysis

Insightful information can be discovered through the knowledge graph. The graph answers basic questions about e.g., *financial analysis, industry analysis, and company analysis.* The following are our competency analysis questions which we aim to answer.

- *Financial analysis e.g., from the balance sheet, profit and loss statements*
    - What is the revenue of companies in 2018 and 2019 and the year-over-year (YOY) growth?
    - What is the gross margin of companies in 2018 and 2019?
    - What is the current ratio of companies in 2018 and 2019?
- *Industry analysis*
    - How many companies have an asset less/more than GBP 10,000 in 2019?
    - What is the profit of companies registered in SIC code 68209[9]?
    - What are the top 5 industries (SIC code) which have the largest average number of employees in 2019?
- *Company analysis*
    - Which companies have parent and/or subsidiary companies?
    - What is the full name of Mr. J D Millet – the director of the company Newgate Kennels Limited? Is there any further information we can learn more about this person e.g., birthday, nationality, occupation, and address?

Section 4.2 describes how these questions are designed in more detail.

Goal 3: Data visualization

Visualizing data is another method to explore our data. In doing this, it can provide a clearer picture and potentially more rigorous analysis for both analysts and consumers. The aim of this paper is to use information from the graph and visualize it in a business intelligence tool such as *Tableau*[10] or an application like *Google Maps.*

[9] Standard Industrial Classification code 68209 – Letting and operating of own or leased real estate

[10] https://www.tableau.com/

For example, the company location data from the graph can be plotted on the map and provide other details e.g., company name and company number, number of employees, business activity as well as useful sources for further information. This interactive map view allows the users to effectively communicate with data.

In addition, an insight can be discovered through the visualization, for instance, *a connection between an officer and companies* or *a relationship between companies*. In this task, the tool called *Tarsier*[11] was used to obtain and visualize further in-depth information on the relationship between officers and companies, if any.

All in all, the findings should make an important contribution to the field of the interoperability of open financial data. Additionally, a practical guideline is provided to encourage creation of a knowledge graph by utilizing corporate internal data or other open data. Data for this study were collected primarily from the CH at various time points during September 2020 and June 2021.

2 General and theoretical background

2.1 iXBRL overview

2.1.1 XBRL vs. iXBRL

XBRL is the open international standard managed by XBRL International for digital business reporting. It is an XML-based format which allows organizations to exchange information rapidly, accurately, and digitally. What makes XBRL reliable is the information tagging, where the reported items are uniquely tagged (An Introduction to XBRL, 2021).

In addition to financial statements, XBRL can also be implemented in other business reports such as *financial information, non-financial information, general ledger transactions, regulatory filings, annual and quarterly reports, risk and performance reports* (The Standard for Reporting, 2018; Extensible Business Reporting Language (XBRL) 2.1, 2013; Roohani, 2008).

The benefit of using XBRL can be seen in the case of a department within the state of Nevada where the controller addressed that XBRL met the goals, such as timely and accurate data, stronger internal controls, reduced costs, and a standardized system of

[11] https://github.com/desmovalvo/tarsier

seamless data exchange (Hoffmann & Rodríguez, 2013). Currently, XBRL use cases are found in more than 50 countries (An Introduction to XBRL, 2021).

Due to the need for financial and business information published in both machine- and human-readable formats, Inline XBRL Specification was extended from the XBRL 2.1 standard. Implemented since 2013, the current version is *Inline XBRL Specification 1.1*[12] which is a second release of the standard (Inline XBRL Part 1: Specification 1.1, 2013). Whereas HTML is used to display data and describe a webpage structure, XML is used for data storage and sharing. Similarly, iXBRL was developed as an alternative form to exchange XBRL data with report presentations. In other words, iXBRL is an HTML document with XBRL tagged data embedded in it.

Unlike XBRL documents, reports in iXBRL format do not need any special tools and can be opened in any web browser. In addition, with the XBRL software, XBRL data in the iXBRL report can be extracted. A direct connection between figures and text in the reporting presentation, and the values of the XBRL facts are key features of iXBRL (iXBRL Tagging Features, 2019; The XBRL Standard, 2021).

Globally, iXBRL is implemented in the following countries (iXBRL, 2021).

- In the UK, over two million companies file iXBRL to HMRC and CH annually.
- In the US, companies can file iXBRL to the SEC, such as 10-Q filing.
- In Denmark, over 100.000 iXBRL financial statements are filed by companies to the Danish Business Registrar.
- In Japan, over 9.000 listed companies and investment funds file iXBRL financial statements to the Japan Financial Services Agency (JFSA).
- In Europe, iXBRL is used as the standard behind ESEF for mandatory IFRS based Annual Financial Statement filings of all public companies.

2.1.2 iXBRL in the UK

Since 1 April 2011, most companies in the UK must file their tax returns, financial accounts, and computations in iXBRL format for the accounting period after 31 March 2010. Additionally, other companies such as overseas companies' resident in the UK,

[12] http://specifications.xbrl.org/work-productindex-inline-xbrl-inline-xbrl-1.1.html

and a company not resident in the UK, but carrying on a trade in the UK through a permanent establishment, branch, or agency in the UK must deliver the required documents in iXBRL format (XBRL guide for business, 2020).

For the taxonomies adopted in the UK, there are three main taxonomies used for company reporting: *Financial Reporting Council (FRC) Taxonomy* set for accounts, *the Corporation Tax or 'CT' Computational Taxonomy* for tax computations, and *the Detailed Profit and Loss taxonomy* for detailed profit and loss statements that are attached to the accounts or computations (XBRL guide for business, 2020). HMRC is responsible for the development of the last two taxonomies. Since 2013, FRC has been responsible for the taxonomy development of financial reporting. FRC is the regulatory body who sets accounting standards in the UK and Republic of Ireland (Company Reporting in the UK – an XBRL Success Story, 2015, p. 5). The XBRL taxonomies published by the FRC reflect the latest version of specifications of XBRL Specification 2.1, XBRL Dimensions Specification 1.0, and Inline XBRL Specification 1.1 (Developer Guide - FRC Taxonomies, 2019, p. 2). To date, there are three versions of the taxonomies which are 2014, 2018, and 2019.

In terms of accounting regulations, the UK began using two major taxonomies (i.e., a UK GAAP taxonomy and a UK IFRS taxonomy) for mandatory filings in 2011. Most companies filed under UK GAAP and publicly quoted organizations used IFRS. These two taxonomies applied different tags to reflect the different requirements under the two regulations (Company Reporting in the UK – an XBRL Success Story, 2015, p. 5).

Regarding the filing of XBRL documents, there are two government bodies involved, namely CH, and HMRC. CH is an executive agency of the Department for Business, Energy & Industrial Strategy, which is a department of Her Majesty's Government (Companies House, 2020).

HMRC is a non-ministerial department, supported by two agencies and public bodies, for the collection of taxes e.g., income tax, environment taxes, customs duty, and excise duties (HM Revenue and Customs, 2021). All registered companies in the UK are required to submit company tax computations and accounts in iXBRL format to HMRC and CH. In October 2010, HMRC and CH introduced a joint filing service for company accounts – a *one-stop* online facility (XBRL guide for businesses, 2020).

iXBRL is used for the voluntary filing of annual reports and accounts to CH. The three main responsibilities of CH are 1) to incorporate and dissolve limited companies, 2) to examine and store company information delivered under the Companies Act and related legislation, and 3) to make information available to the public (Companies House, 2020).

According to the official statistics at CH, there were 4.3 million companies and corporate bodies registered at CH at the end of March 2020. This showed an increase of 3.5 percent from the end of March 2019. In terms of corporate body types, the three corporate bodies account for over 98 percent of all 28 corporate body types. These three types are private limited companies, limited partnerships, and limited liability partnerships (LLPs) (Companies register activities: 2019 to 2020, 2020).

2.1.3 iXBRL architecture

As discussed above, iXBRL is an alternative method to allow the preparer of XBRL instances to specify the desired presentation of the report. It uses the HTML standard to visualize the XBRL data. By embedding additional tags into the HTML code, the XBRL structure and data remains machine-readable and human-readable in the form of an HTML report. In simple terms, both the presentation of accounts and delivery of XBRL data are processed in a single iXBRL file (Company Reporting in the UK – an XBRL Success Story, 2015, p. 4).

Regarding the iXBRL instance's elements, data values are nested within iXBRL Elements which are themselves nested within HTML elements ("Markup Elements").

In this way, the value of each XBRL fact can be displayed by a browser (Inline XBRL Part 1: Specification 1.1, 2013). Figure 2 shows iXBRL components as well as relationships between an iXBRL instance and a taxonomy.

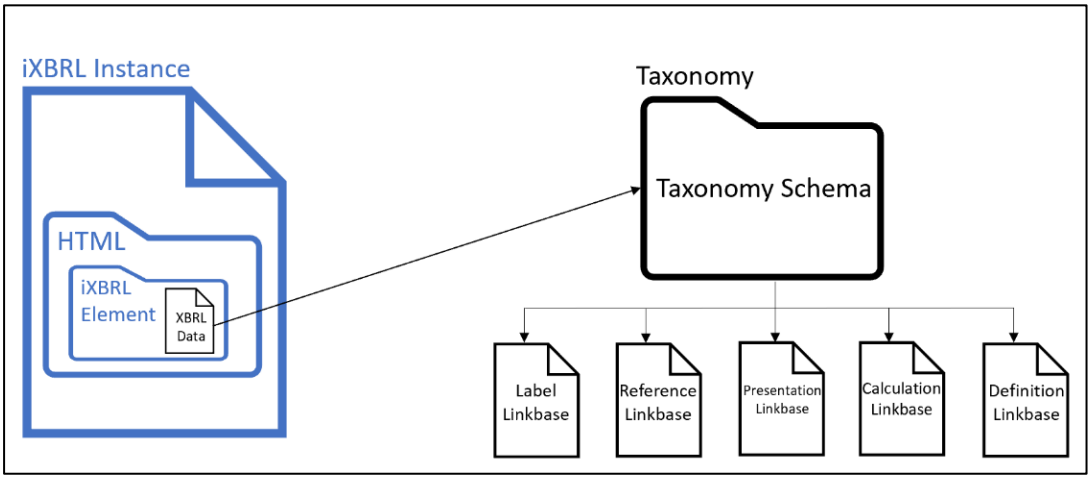


Figure 2: An overview of iXBRL architecture (Inspired from Inline XBRL Part 1: Specification 1.1 (2013) & Extensible Business Reporting Language (XBRL) 2.1 (2013))

XBRL data contains *reported data or facts* which are represented by elements in an instance document, and *concepts* which correspond to element definitions in *an XML Schema or a Taxonomy Schema*. A *taxonomy* contains business reporting vocabularies of reporting concepts that are used in instance reports. It is made up of one (or more) *XML Schema files* to capture terms and presentation grouping; and *Linkbases* which are broad-based structures for capturing *inter-concept relationships* (i.e., Presentation, Calculation, and Definition), *relationships between concepts and their documentation* (i.e., Labels and Reference) (Extensible Business Reporting Language (XBRL) 2.1, 2013). The following parts describe each component, namely iXBRL instance and XBRL taxonomy in more detail.

2.1.3.1 iXBRL instance

An instance document is a collection of facts which are described by value, concept, and context. For example, 100 is the value of the fact with the concept ‘Revenue’ where the context explains the entity, period, and currency of the revenue reported (Getting started for Developers, 2021). As mentioned earlier, an iXBRL instance is a well-formed XML document. It contains both *Markup Elements* and *iXBRL Elements*. A processor will ignore the Markup Elements and combine the iXBRL Elements and the data values to generate an XBRL instance document (Inline XBRL Part 0: Primer 1.1., 2015). Thus, a number appearing in the iXBRL report is not necessarily identical in format to the underlying XBRL value (iXBRL Tagging Features, 2019).

Referred to Inline XBRL Specification 1.1, an iXBRL Element is any element with an 'ix' namespace name with a value of `http://www.xbrl.org/2013/inlineXBRL`. Examples of iXBRL Elements include the following. More details on iXBRL elements are available in the full specification.

- The `ix:header` element contains the non-displayed portions of the instance such as `ix:hidden`, `ix:references`, and `ix:resources`.
- The `ix:hidden` element is used to contain XBRL facts of which are not expected to be displayed in the report e.g., a repeated value which is required to display once.
- The `ix:continuation` element is used to define data that is to be treated as part of `ix:footnote` or `ix:nonNumeric` elements.
- The `ix:nonNumeric` element denotes an XBRL non-numeric item.
- The `ix:nonFraction` element denotes an XBRL numeric item.

Figure 3 shows an example how a nonFraction fact is displayed in a financial report.

EDWARD BAARDA LIMITED (REGISTERED NUMBER: 00374596)

NOTES TO THE FINANCIAL STATEMENTS - continued
for the Year Ended 31 December 2019

4. **TANGIBLE FIXED ASSETS**

	Freehold land £	Greenhouses and buildings £	Plant and machinery £	Motor vehicles £	Totals £
COST					
At 1 January 2019	243,446	1,661,119	3,210,982	184,795	5,300,342
Additions	-	-	4,608	-	4,608
At 31 December 2019	243,446	1,661,119	3,215,590	184,795	5,304,950

Figure 3: Example of a nonFraction fact in a financial report

2.1.3.2 XBRL taxonomy

As explained earlier, the reporting fact is described by the concept which relates to the taxonomy. The taxonomy allows the author to report data by using an XML Schema document which includes element definitions, and a collection of extended links (Linkbases) that forms part of the concept definitions (Inline XBRL Part 0: Primer 1.1., 2015). The Linkbases in a taxonomy describe the meaning of the concepts by expressing relationships between concepts and by relating them to their documentation (Extensible Business Reporting Language (XBRL) 2.1, 2013).

Without the taxonomy, an XBRL instance would lack meaning (XBRL Taxonomy Development Handbook, 2020, p.29).

The purpose of XML Schema files is to collect terms and presentation groupings as well as to define the actual concepts forming as the basis of taxonomy. It stores the name, data type, period type, and how they might be used. An XBRL Linkbase file, on the other hand, includes the explicit relationship definitions between concepts defined in the XBRL Schema (Cohen, 2012).

There are *five different kinds of Linkbases* used in taxonomies to document concepts: definition, calculation, presentation, label, and reference. While Definition Linkbase, Calculation Linkbase and Presentation Linkbase express inter-concept relationships, Label Linkbase, and Reference Linkbase express relationships between concept and their documentation (Extensible Business Reporting Language (XBRL) 2.1, 2013).

In terms of the taxonomy structure, one taxonomy comprises different files with links to each other. The main taxonomy file which contains a basic list of tags is a schema file with an `.xsd` filename extension.

Other files including lists of labels, the presentation view, references, dimensional definitions, and other information are Linkbases with `.xml` extensions. There are also other files covering standard entity and business data, the directors' report, audit report, and an accountant's report (XBRL Tagging Guide –FRC Taxonomies, 2019, p. 40).

2.2 Semantic web

Semantic web technologies such as linked data, vocabularies, inference, query languages are developed for people to use as tools to create data stores on the web, build vocabularies, and write rules for handling data (Semantic Web, 2019). By implementing the technologies, people can discover and reuse data in interesting or unexpected contexts (Hall & O'Hara, 2009, p. 8085). The goal of the *Web of Data* is to allow machines to exchange content with each other in a structured format without the need for human intervention or guidance (Hogan, 2020, p. 15; Ashraf & Hussain, 2012; Bao et al., 2010).

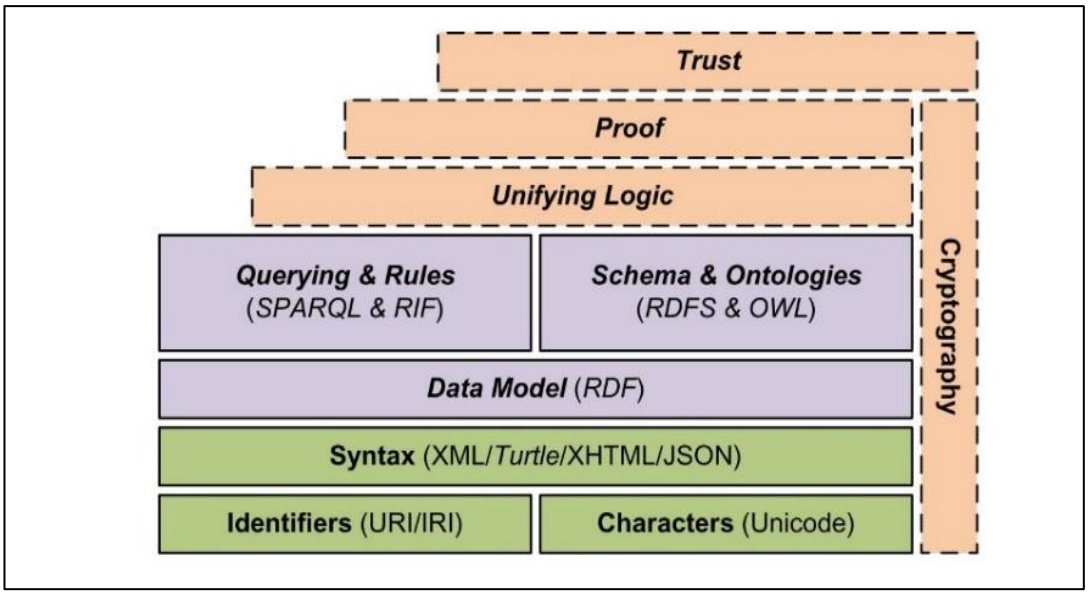


Figure 4: Semantic web stack

Figure 4 depicts the semantic web stack. Starting from the bottom of the stack or the foundation of the semantic web, this part borrows directly from existing web technologies. To be able to map from binary sources and storage to textual information, the standard unicode characters are required. Similar to the identification of documents on the web, the *Uniform Resource Identifier (URI)* is used to identify things in a machine-readable manner for the semantic web. As URIs only support a subset of American Standard Code for Information Interchange (ASCII) characters, Internationalized Resource Identifier (IRI) was proposed to cover a larger set of characters (Hogan, 2020; Hogan, 2014;Dürst und Suignard, 2005). In addition to existing generic syntaxes (e.g., XML and JSON), the Terse RDF Triple Language (Turtle) syntax was created to encode semantic web data. This syntax allows machines to parse data automatically.

The middle part is where the heart of the semantic web is located. Machines can exchange data, understand the meaning of content, and generate results through tools like data-model, schema & ontologies, and querying & rules. Through a number of standardization efforts, RDF, RDF Schema (RDFS), Web Ontology Language (OWL), SPARQL Protocol and RDF Query Language (SPARQL), and Rule Interchange Format (RIF) have been developed and realized.

While a core data model of RDF is used on the semantic web, RDFS and OWL are used to bring semantics to describe things in RDF content. To query the RDF content, SPARQL is the current querying standard for the semantic web.

In addition, RIF is the rule standard used to capture the expressivity of various existing rule-based languages (Hogan, 2014; Wang et al., 2014).

The semantic web is to put and link data on the web. The links are between resources or things on the web which can be described by using RDF. In order to publish data on the web, the four Linked Data principles are as follows:

1. *use URIs as names for things*
2. *use HTTP URIs so that people can look up those names*
3. *when someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)*
4. *include links to other URIs, so that they can discover more things*

– described by Tim Berners-Lee (2009).

2.2.1 RDF

Resource Description Framework (RDF) was recommended by W3C in 1999, with the most recent version – RDF 1.1 released in 2014. It is a standard representation of information on the web. Unlike other formats in a table- and tree-structure (e.g., CSV, JSON, and XML), RDF is structured as a graph-based data model that consists of a subject, a predicate, and an object. This triple set is therefore called an RDF graph. When modeling the data in RDF format, one can represent such information in a flexible way (Hogan, 2020, p. 60; Resource Description Framework (RDF): Concepts and Abstract Syntax, 2004).

By using RDF, we can describe resources such as virtual entities (e.g., web pages and websites), concrete entities (e.g., people and places), and abstract entities (e.g., categories, ancestry relationships, and points in time). In simple words, resources are anything with an identity that one could consider describing data. There are three types of RDF terms that can be used to describe the resources, namely IRIs, literals, and blank nodes (Hogan, 2020).

As was mentioned in the previous part, we can use URIs or IRIs to identify resources and properties. An RDF URI reference is a Unicode string that does not contain any control characters (i.e., non-printing characters) such as #x00 (null characters) (Resource Description Framework (RDF): Concepts and Abstract Syntax, 2004). As described in RDF 1.1 Concepts and Abstract Syntax (2014), for non-ASCII characters,

the mapping involves UTF-8 encoding and %-escaping octets that are not allowed in URIs.

In case a based IRI is specified and can be resolved with relative IRIs to make them absolute, the usage of relative IRIs for such absolute IRI references is allowed as a shorthand (Resource Description Framework (RDF): Concepts and Abstract Syntax, 2004).

Furthermore, the IRI reference may include a fragment identifier to identify a secondary resource that is usually a part of, view of, defined in, or described in the primary resource. For example, `eg:someurl#frag` is the thing indicated in the document at `eg:someurl`. IRIs in RDF graphs can indicate anything such as something external to the representation, or external to the web. Thus, the RDF-bearing representation can be considered as an intermediary between the web-accessible primary resource, and some set of non-web or abstract entities described by the RDF graph (RDF 1.1 Concepts and Abstract Syntax, 2014). Importantly, as stated in the Linked Data principle, HTTP IRIs should be used.

While the IRI reference is used to describe resources, literals are used to identify values including strings, numbers, booleans, and dates. A literal in an RDF graph consists of two or three elements, namely a lexical form, a data type IRI, and a language tag. Examples of literals are `"Nuremberg"@en`, `"1"^^xsd:integer`, `"true"^^xsd:boolean`, and `"2021"^^xsd:gYear`.

The data type abstraction used in RDF is compatible with XML Schema (RDF 1.1 Concepts and Abstract Syntax, 2014).

In terms of the blank node, it can be used to represent a resource for which an IRI is not given or unknown. To represent a blank node, one can use an underscore prefix, for example, `_:bn1`. Instead of having a global scope like a URI or IRI, a blank node is treated as a variable with a local scope. The use of a blank node identification declares the existence of a certain resource (Hogan, 2020, p. 70).

To express a statement in an RDF graph or a set of triples, it consists of three components: a subject, an object, and a predicate. Each triple represents a statement of a relationship between things. This relationship can be illustrated by a node-arc-node link with the arc always pointing to the object. Depending on the data, a node can be in any form of RDF terms, namely a URI or IRI, a literal, or a blank node.

However, there are some restrictions on where different types of RDF terms can be placed (Resource Description Framework (RDF): Concepts and Abstract Syntax, 2004):

- Subject position: IRIs or blank nodes;
- Predicate position: IRIs only; and
- Object position: IRIs, blank nodes, and literals.

To serialize RDF in files for storage, parsing, and processing, there are currently various RDF syntaxes proposed such as RDF/XML (an XML-based syntax), N-Triples (simple line-based syntax for RDF), Turtle (human-friendly syntax for RDF), RDFa (an embedded syntax in HTML), and JSON-LD (a JSON-base RDF syntax) (Hogan, 2020, p. 106).

Having defined the relevant RDF terms and the graph model, construction of RDF statements is shown in the following example. This can be illustrated briefly by using such concepts to describe the below statements about Max Mustermann. In this case, the URI of `http://example.org/` with `ex:` prefix is used.

- maxMustermann is a person and is a student.
- maxMustermann is called "Max Mustermann" in German.
- maxMustermann is an author of a paper.
- maxMustermann is studying at FAU located at Lange Gasse 20, 90403 Nuremberg.

The first three statements are considered simple statements where we can identify subject, predicate, and object. It is clear that the subject for such statements is the same resource – Max Mustermann, and it can be represented as the URI of `http://example.org/maxMustermann` or `ex:maxMustermann`. Depending on our needs, the first statement can be modeled by using `ex:isa` with a person entity and a student entity; or creating a new predicate and using literal with data type boolean.

The second statement is where we can use `ex:hasName` property to describe the resource's name with the language tag of German. Again, we can create a URI predicate to indicate the relationship of the author as `ex:isAuthor` and an object of paper resource as `ex:aKGfromUKFinancial Statements`.

The corresponding RDF statements are presented below. By presenting in a Turtle syntax, two shortcuts can be used, namely a semicolon "`;`" for triples that describe the same subject, and a comma "`,`" for another object with the same subject and predicate (Berners-Lee, 2006).

In order to model a complex relation – a relationship between more than two resources, an approach – *n-ary relation modelling* can be used (Hogan, 2020, p. 79). For the last statement, one could represent it in RDF as follows:

```
ex:maxMustermann ex:studyAt ex:FAU.

ex:FAU ex:address ex:FAUaddress.

ex:FAUaddress ex:street "Lang Gasse 20"^^xsd:string.
```

The URI `ex:FAUaddress` identifies an abstract resource representing the FAU address. Such a resource is an n-ary relation. In addition, the n-ary relation can be represented by using the blank node `_:bn1`.

2.2.2 Ontologies

After structuring data in the RDF model, ontologies or vocabularies can be added to the RDF data as a description of such concepts or relationships. In doing this, it can help to reduce the ambiguities existing in the data sets or can help to discover new relationships (Ontologies, 2021).

In addition, the ability to capture meaningful generalizations about data in the web can be enhanced by using richer vocabularies or ontologies such as Web Ontology Language (OWL), inference rule languages, and other formalisms (e.g., temporal logics) (RDF Schema 1.1, 2014).

On the web, two main standards, namely RDF Schema (RDFS) and OWL are proposed to define semantics over RDF data. RDFS is an extension of the basic RDF vocabulary. This data-modelling vocabulary for RDF data enables the expression of the semantic definitions needed to automate the deductions.

The latest standard – RDFS 1.1 was released in 2014. To denote instances of classes, one can use RDF which provides a built-in property such as `rdf:type`. Examples of RDFS key terms to specify relationships between classes and properties are

rdfs:subClassOf, rdfs:subPropertyOf, rdfs:domain, and rdfs:range.

While RDF defines the meta class rdf:Property as the class of all properties, four new meta classes are defined by RDFS: rdfs:Resource, rdfs:Literal, rdfs:Datatype, and rdfs:Class. In addition, rdfs:label, rdfs:comment, rdfs:seeAlso, and rdfs:isDefinedby can be used to annotate resources (Hogan, 2020).

Based on the example above, the predicate ex:isa can be changed to rdf:type in order to indicate that ex:maxMustermann is an instance of a class person. In addition, the predicate ex:hasName can be changed to rdfs:label to provide a human-readable format of the resource's name. For the predicate ex:isAuthor, we can state its domain and range as

```
ex:isAuthor rdfs:domain ex:Person;
            rdfs:range ex:Report.
```

From these two claims, it can then entail the below statements.

```
ex:maxMustermann rdf:type ex:Person.
ex:aKGfromUKFinancialStatements rdf:type ex:Report.
```

The Linked Data best practices suggest the reuse of existing vocabulary terms across the web instead of developing new vocabulary terms again and again. The interoperability of the corresponding data can significantly increase if the reused vocabularies are deployed across different websites (Hogan, 2020. The following is a list of existing ontologies developed to describe various resources in different domains (Roman et al., 2021):

- For organizational structures: the W3C Organization Ontology (ORG), the e-Government Core Vocabularies, the Registered Organization Vocabulary (RegOrg), and Schema.org.
- For financial and economic use cases: the Financial Industry Business Ontology (FIBO), the Entity Legal Forms Code List, and Wikipedia's Legal Entity Types.

- For company identification and location: the Global Legal Entity Identifier Foundation (GLEI), the Business Registers Interconnection System (BRIS), the Territorial Units for Statistics (NUTS), Local Administrative Units (LAU), the GeoVocab.org, and GeoNames.
- For other fields: Friend of a Friend (FOAF), Dublin Core, DBpedia, Bibliographic Ontology (BIBO), ADMS, Vocabulary of Interlinked Datasets (VoID), IANA language code registry, Person Core Vocabulary, and the Simple Event Model Ontology (SEM).

To apply the vocabularies in our example, we obtain external vocabularies from Schema,[13] FOAF,[14] BIBO,[15] DBpedia,[16] and vcard.[17]

In the case that there is no existing vocabulary suitable to describe our data, a new vocabulary should be developed (Hogan, 2020, p. 540). In this case, we define `ex:isStudent` by using RDF, RDFS, OWL vocabularies and link our resources to the external vocabulary as well.

Based on the property defined, we can entail from `rdf:type`, `rdfs:domain`, `rdfs:range`, and `owl:DatatypeProperty` that the subject of the statement is an instance of a class from `foaf:Person` and the object should have the data type of boolean.

In addition, external URIs can be dereferenced to get further information (Hogan, 2020, 537). Therefore, we use the Wikidata IRIs to allow a client to dereference additional useful data about the FAU University –Q40025 and Nuremberg – Q2090. Based on these semantic web and existing external vocabularies as well as Wikidata entities, a graph could provide richer semantics.

2.2.3 Query languages

For data in RDF format, SPARQL can be used to query such data for the semantic web. SPARQL become a W3C recommendation in 2008 (Hall & O'Hara, 2009, p.

[13] http://schema.org/

[14] http://xmlns.com/foaf/0.1/

[15] http://purl.org/ontology/bibo/

[16] http://dbpedia.org/ontology/

[17] http://www.w3.org/2006/vcard/ns#

8085). The SPARQL 1.1 standard[18], finalized in 2013, provides languages and protocols to query and manipulate RDF graph content on the web or in an RDF store. Extended from an earlier standard in 2008,[19] subqueries, value assignment, path expressions, and features of aggregates (e.g., `COUNT`) have been included in the 2013 standard – SPARQL 1.1 (SPARQL 1.1 Overview, 2013).

Starting with four simple queries: `SELECT`, `DESCRIBE`, `CONSTRUCT`, and `ASK` queries are described below.

The `SELECT` query is for the solution of the number of humans in Wikidata database. The `COUNT` aggregate feature is used to provide the number. The result shows over nine million entities which are instances of the human entity (i.e., wd:Q5 or <www.wikidata.org/entity/Q5>). We note that every entity in Wikidata is associated with a unique Q-code and every property in Wikidata is associated with a unique P-code.

To identify the resources, a simple `DESCRIBE` query can be used. Similar to the `DESCRIBE` query, a `CONSTRUCT` query returns an RDF graph constructed by a graph template.

2.3 RML

RML or RDF Mapping Language is a generic RDF mapping language. It was the first language proposed to extend from the Relational to RDF Mapping Language (R2RML), recommended by W3C in 2012 (W3C, 2021; Janev et al., 2020, p. 60). While R2RML is used in mappings from relational databases or tabular structures to RDF representation (R2RML: RDB to RDF Mapping Language, 2012), RML is a superset of R2RML and is defined to facilitate one or more data sources in either similar or different formats.

Regarding the sources and formats, RML supports other data sources than the relational databases such as CSV, XML, JSON, and access interface, e.g., files or web APIs (Heyvaert et al., 2019). Therefore, RML is proposed as an alternative solution to cope with the heterogeneous data issue (Dimou et al., 2014).

[18] https://www.w3.org/TR/sparql11-overview/

[19] https://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/

3 Architecture and method of the knowledge graph

3.1 Architecture of the knowledge graph

To enable a better understanding of the guideline, a framework was developed to propose an underlying architecture of the knowledge graph construction. In doing this, Figure 5 borrows the framework from (O'Riain et al., 2012). This model is a good illustration of open data source integration using linked data. Especially as the comparability of the data is important (Wenger et al., 2013). The design of our model below was slightly rearranged and tailored from the original figure, following the paper's purposes. In addition, the *Linked Data Cookbook* (2014) – *The 7 Best Practices for Producing Linked Data*[20] provided a basis throughout the steps proposed. The figure represents a processing architecture as well as the relationship of an information value chain, an information architecture, and the steps in creating the knowledge graph. The numbers in yellow denote the steps.

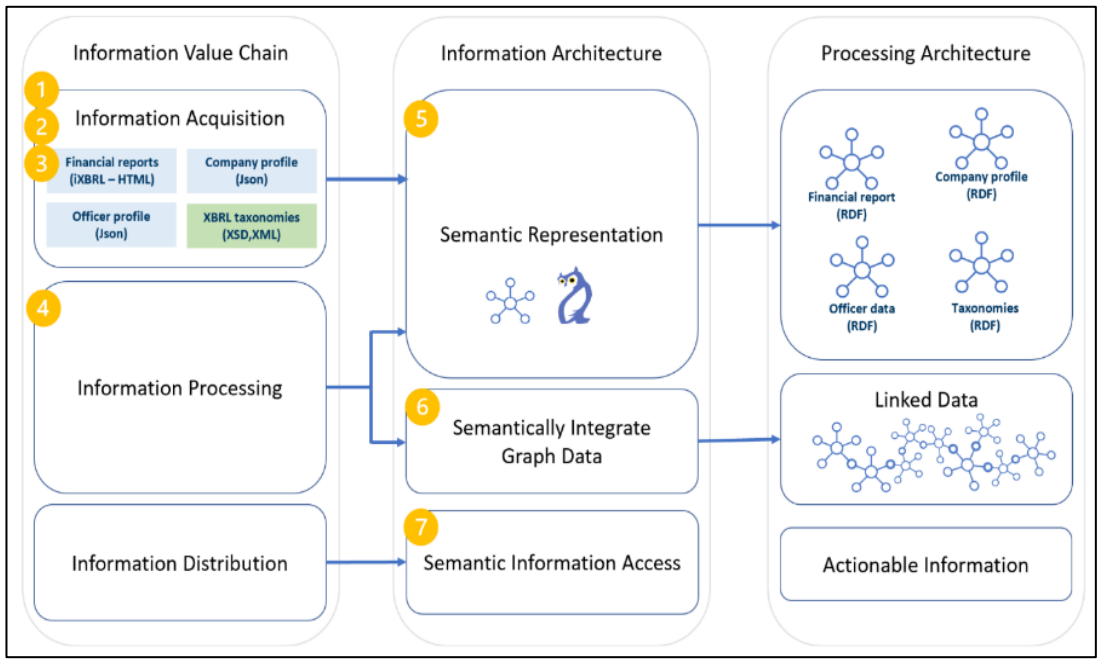


Figure 5: Framework of data integration and steps to create the graph.
Adapted from O'Riain et al. (2012)

Starting from the first pillar, the information value chain comprises three value-adding activities, namely information acquisition, information processing, and information distribution.

[20] https://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook

Information Acquisition includes sourcing, gathering, and preprocessing data from different sources. Hence, it consists of three steps in our guideline: **Step 1: Define goal and data requirement of the graph**, **Step 2: Gather the relevant data,** and **Step 3: Preprocess the data**. After these steps, the data is ready to be transformed into the RDF representation – **Step 5: Transform the data to RDF**. This step results in a generation of an individual graph from each source.

In the meantime, activities that support the data integration are performed under *Information Processing.* To transform the data by using RML, mapping rules need to be specified, including the data model.

Therefore, the step under this activity is **Step 4: Model RDF data**. After the data transformation in Step 4 and Step 5, different graph data is integrated in **Step 6: Link entities**. For example, the graphs of an iXBRL report, XBRL taxonomies, company profile, and officer profile are combined into one graph. The semantic representation is hence in RDF format and embedded with RDFS, OWL[21], and existing ontologies (OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax, 2012). To give an overview from the short review above, Figure 6 summarizes all information in our knowledge graph.

Finally, *Information Distribution* provides the user access to the semantic information to further utilize in other tasks e.g., prediction, visualization, or modelling. The linked data is stored in the triple store, which in our case is Apache Jena Fuseki,[22] so users can query the data by using the SPARQL query language. This activity is performed under Step 7: Evaluate results, where our goals of graph quality, data analysis, and data visualization use cases are included in Section 4 – Evaluation.

[21] https://www.w3.org/OWL/

[22] https://jena.apache.org/documentation/fuseki2/index.html

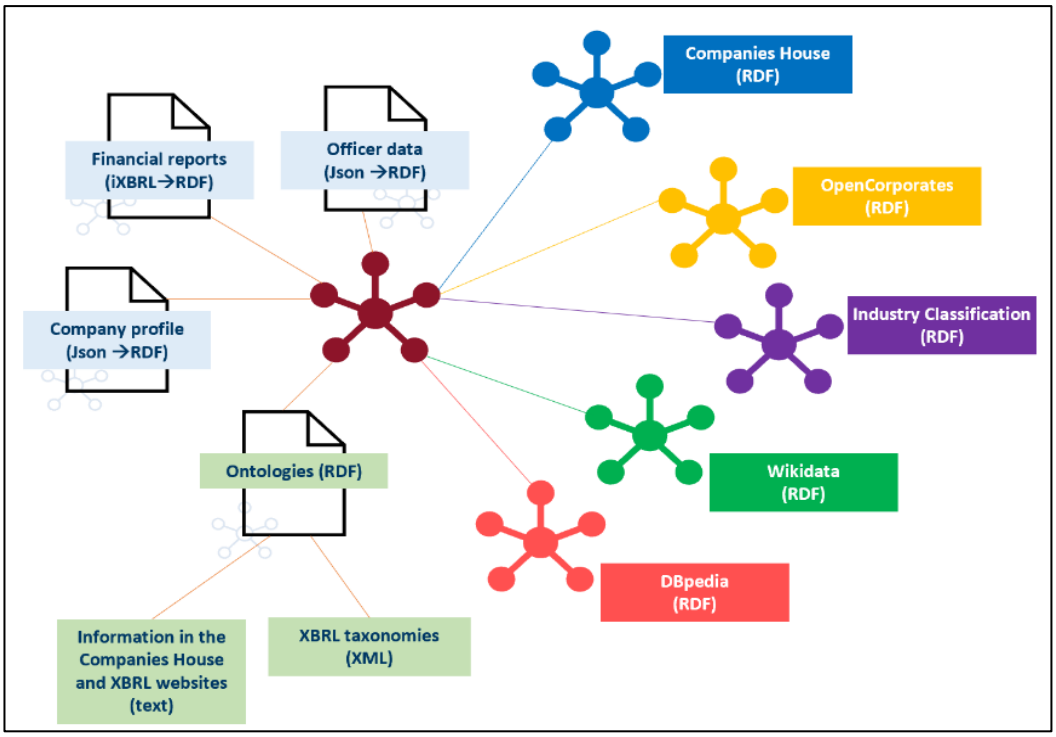


Figure 6: Knowledge graph components

3.2 A step-by-step guideline

3.2.1 Step 1: Define goal and data requirement of the graph

Table 2 summarizes the information required to analyze the underlying research questions.

Competency questions	Information required	Sources	Data format
What is the revenue, YOY growth, gross margin, and current ratio of companies in 2018 and 2019?	iXBRL financial reports	CH – bulk download and XBRL taxonomy	HTML and XML
How many companies have an asset less and more than GBP 10,000 in 2019?			
What is the profit of companies registered in SIC code 68209?	iXBRL financial reports, company profile, and industry classification	CH – bulk download, Web API, and industry code	HTML, JSON, and RDF
What are the top 5 industries (SIC code) which have the largest average			

Competency questions	Information required	Sources	Data format
number of employees in 2019?			
Which companies have parent and/or subsidiary companies?	Company profile	Web API	JSON
What is the full name of Mr. J D Millet – the director of the company Newgate Kennels Limited? Is there any further information that we can learn more about this person e.g., birthday, nationality, occupation, and address?	iXBRL financial reports and officer profile	CH – bulk download and Web API	HTML and JSON

Table 2: List of information requirement by the competency questions

To answer the competency questions, information of financial figures, company and officer profiles, and industry classification are required. This information is however in different formats. Therefore, the common data representation is useful for the data integration task, such as RDF format (O'Riain et al., 2012). External graphs are also included in this work, particularly Open Corporates, DBpedia, and Wikidata. We note that the questions are still answered without these graphs. However, due to our intention to create sustainable development, connecting to other ongoing developed graphs will be beneficial to our knowledge graph.

3.2.2 Step 2: Gather the relevant data

To summarize, there are five sources, from which the data can be obtained: 1) CH (i.e., bulk data, web API, and Linked Data service), 2) FRC XBRL taxonomy, 3) Open Corporates, 4) DBpedia, and 5) Wikidata.

For the data in other formats than RDF (i.e., from CH and the taxonomy), the transformation into the RDF format is necessary. On the other hand, the data in RDF format (i.e., from the external graphs) can be integrated by entity linking. While the entity linking is explained in Step 6: Link entities, a collection of the RDF data is detailed under this step. The following is a detail of data collection by sources.

In total, we collect 1.149 iXBRL reports, 2.846 company profiles, and 5.293 officer appointments (4,331 officers) from CH. To resolve the different years of taxonomy issue (e.g., an update of taxonomy from the previous year), 2014 and 2019 FRC taxonomies arechosen based on most of our collected financial reports. For the entity linking, we collect 421,432 company entities and 743,584 person entities from DBpedia.

3.2.2.1 Data source – CH

The data that we can collect from CH are iXBRL reports, company profiles, officer profiles, and industry classification. While it is possible to simply collect iXBRL reports and industry classification in a traditional manner (i.e., download via clickable links), company and officer profiles are available through web API.

For the company profile, not only general company information is disclosed, but also the company's account and branch details, if any. The data type of the information varies, such as integer, string, and boolean. On the other hand, the officer profile information is based on the officer's appointment. The same officer may be appointed several times with same/different roles/companies in which his/her personal information might change throughout the period e.g., officer role, address, occupation, and nationality. Thus, the officer link, in this case, refers to an appointment instead of an individual officer.

In addition, the linked data of the companies can be queried via CH SPARQL Endpoint[23]. The query result is available in multiple formats including table, JSON, and XML. The `DESCRIBE` query provides the information of the company Zenith Print (UK) Limited such as company number, business activity, as well as incorporation date. As shown in the figure, CH reuses well-known vocabularies of skos, org, and rov; and creates its own vocabulary.[24]

[23] http://business.data.gov.uk/companies/docs/getting-started-with-query.html

[24] http://business.data.gov.uk/companies/def/terms/

The vocabularies created by CH are for example: *account category, company category, company status, and industry classification.*[25] To denote the geographical information of the UK company, the postcode ontology from Ordnance Survey[26] is imported (Pan & Zhang, 2016).

As mentioned in Section 1.1.3 – Companies House, CH has defined URIs with company numbers to identify an entity of a company, for example in which we can use this identifier to link with our data.

Regarding the industry classification, *Standard Industrial Classification (SIC)*, a five-digit classification, is used to classify business establishments and other statistical units by economic activity (UK SIC 2007, 2022). To run a business, a company's nature of business is to be provided by using SIC codes. In the case of various products offered or services rendered, a company is allowed to choose up to four SIC codes. The codes can be changed later when the business changes its operating activities (Townley, 2018). CH uses and accepts the filings that contain only SIC codes on the condensed list available from Office for National Statistics (ONS) (Standard industrial classification of economic activities (SIC), 2008).

The most important resources are the SIC codes and their labels. In case a company registers under *SIC 36000*, the main business of such company is therefore in the area of *water collection, treatment, and supply*. And *a retailer of leather goods* would register under *SIC 47720*.

3.2.2.2 Data source – XBRL taxonomy

As explained in Section 2.1.3.1, the main taxonomy file is a *schema file* (XSD format) while the *Linkbases* files consist of labels, presentation views, references, and definitions (XML format). Examples of the taxonomies are Business taxonomy, Countries and Regions taxonomy, and Core Financial Reporting taxonomy. The structure of full folders and files comprising the FRC taxonomy is provided in Appendix A – FRC taxonomy file structure.

[25] http://business.data.gov.uk/companies/def/sic-2007/

[26] http://data.ordnancesurvey.co.uk/ontology/postcode/

3.2.2.3 Data source – Open Corporates

Open Corporates[27] is considered the largest open data collection for companies, covering more than 1 million companies in its database. Specifically for UK-based information, there are around 13 million companies and 44 million officers contained in the database. For the data format, the company data is available in RDF, XML, and JSON format, while the officer data is available in XML and JSON format only.

We note that Open Corporates has defined URIs with company numbers to identify companies, for example, `http://opencorporates.com/companies/gb/02050399`, in which we can use this identifier to link with our data.

3.2.2.4 Data source – DBpedia

The second external knowledge base linked with our graph is DBpedia. The required information to perform the entity linking is company names and person names. Therefore, the English label of the instances of the classes of *Organization (`dbo:Organisation`)* and *Person (`dbo:Person`)*, as well as their subclasses, are obtained. To limit the number of person entities, the property of birth year (`dbo:birthDate`) between 1900 and 2000 is also applied when collecting the data.

In total, we collect 421.432 organization entities and 743.584 person entities from DBpedia.

3.2.2.5 Data source – Wikidata

In a similar way to obtain the data from DBpedia, the English label of the instances of the classes *Company (Q783794)*, *Business (Q4830453)*, and *Human (Q5)*, as well as their subclasses, can be collected. Despite experimenting various approaches, we were unable to gather such a large dataset due to a lack of computational power.

3.2.3 Step 3: Preprocess the data

Data preprocessing is the first step for most of the big data related tasks to understand how healthy or clean our data and its structure is. By collecting data from different sources, the uncertainty of data is an issue that users may encounter; for example,

[27] https://opencorporates.com/

noise, incompleteness, and inconsistency found in the data (Janev et al., 2020, p. 41). Therefore, it can impact the reliability of such collected data and potentially in the workflow and the analytics step (Janev et al., 2020, p. 16). Depending on the dataset, multiple techniques to data preprocessing may be necessary.

Prior to the RDF transformation task, we find that the preprocessing task for the iXBRL report and the schema of taxonomy is necessary. According to the information in iXBRL reports and schema files, a variety of namespaces is implemented. In addition, those namespaces contain the taxonomy information, namely year and category. Referring to the RDF transforming task, a concept of a fact is transformed into RDF format as a URI. However, the concept's URI may be meaningless without the namespace replacement.

3.2.4 Step 4: Model RDF data

Prior to the RDF transformation, we firstly identified a logical model of our datasets, used URIs to name objects, and reused vocabulary, as suggested in the Linked Data Cookbook (Linked Data Cookbook, 2014).

Based on the study of (Radzimski et al., 2014, Mora-Rodriguez, 2017; Asimadi et al., 2017, Sánchez-Cervantes et al. 2018), we adopt the semantic model for financial data and adaptit to fit our paper purpose as shown in Figure 7. The figure depicts a high-level overview of the resulting data model. It highlights the main classes in the model, the relations between them as well as the data types.

This semantic model favors our paper goals and supports data navigation.

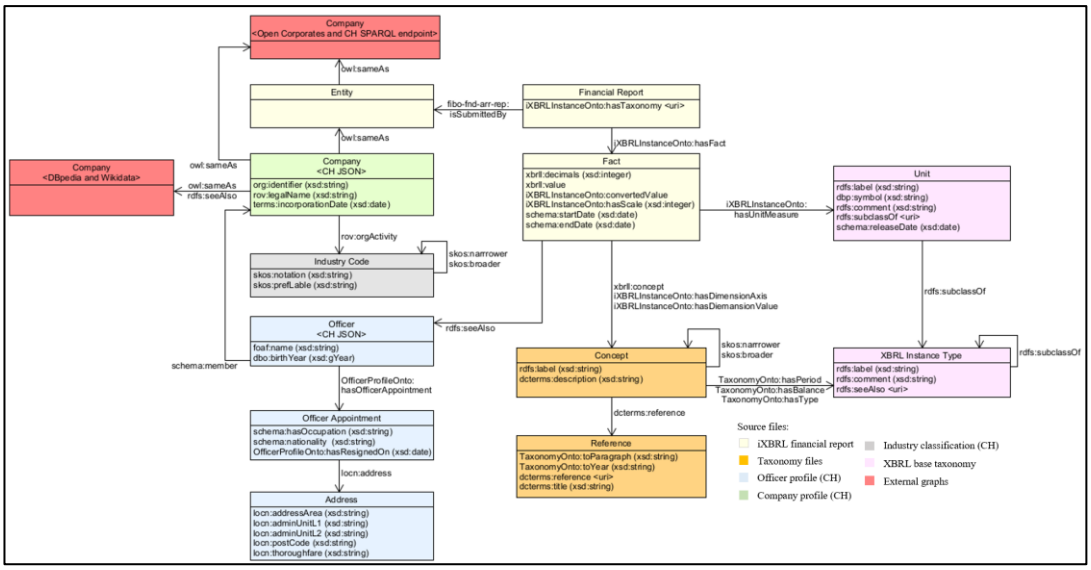


Figure 7: High-level of knowledge graph semantic model

Each model component is discussed in detail in the sections below:

- **Financial report:** An iXBRL financial report contains facts which can be discovered via a relation `iXBRLInstanceOnto:hasFact`. Additional context information, such as value, period, unit, dimension, and concept is derived from a fact. A concept and a dimension of a fact refer to a taxonomy. An entity is the one who submits the financial report and is referred to as the company entity in the external graphs (red boxes) and CH (green box). Where a fact relates to the concept of an officer, such fact refers to an officer entity obtained from CH (blue box) via `rdfs:seeAlso`.
- **XBRL taxonomy:** This component could be considered as an iXBRL instance's ontology, which can provide additional information to users. The taxonomies are structured in XML schemas and linkbases that define the financial report's structure through the following linkbases: label, reference, and presentation. Further detail is provided in Section 2.1.3.1.
- **XBRL base taxonomy:** A unit of measurement is allocated to all values reported. The unit taxonomy contains information of the measuring units used in financial statements.

  It represents information of the financial statements' measurement units (e.g., dollars, shares, euros, or dollars per share) (Sánchez-Cervantes, 2018). Additional information of XBRL instance types (pink box) such as string, monetary, period (duration and instant), and balance (credit and debit) are included in the model as well.

We note that this paper does not include the study of XBRL base taxonomy. While the base taxonomy is not the focus of this study, for completeness we included it in this stage.

- **Officer:** Linking from the fact, more information on the officer can be found, such as the person's full name, birthday, appointment, and address. In addition, a predicate `schema:member` is used to describe a relationship between an officer and a company.
- **Company:** Identical company entities in the financial report, CH, and external graphs are connected via either `owl:sameAs` or `rdfs:seeAlso` (more detail in Section 3.2.6). Business activity of the company can be discovered via `rov:orgActivity` to the industry classification (gray box) in RDF format provided by CH.

3.2.4.1 Ontology

In addition to the RDFS and OWL standards, other linked open vocabularies are used to define our datasets. In doing this, it can reduce diversity and make data integration simpler (Hogan et al., 2020, Livieri et al., 2014, Noy and McGuinness, 2001). The vocabularies adopted in our paper are such as xbrll, dcterms, foaf, skos, schema, gleig, rov, and ebg. The full list of vocabularies with their namespaces can be found in the data model.

Where we can't find the applicable vocabularies, we develop our own ontologies. Our ontologies are classified based on the data which are 1) iXBRL ontology, 2) taxonomy ontology, 3) officer ontology, and 4) company ontology.

In developing such ontologies, Protégé[28] is a suitable tool for this. Protégé is a free and open-source platform that can be used to construct ontologies.

In addition to linked data vocabularies and our own ontologies, XBRL taxonomy which is explained in the following part can also provide semantics to the financial report.

[28] https://protege.stanford.edu/

3.2.5 Step 5: Transform the data to RDF

Following the structure of our data model, the data in different formats are lifted into the common format – RDF in this step. To demonstrate how we transform the data into RDF, we provide the following examples of HTML format – iXBRL report, XML format – taxonomy, and JSON format – company and officer profiles. As previously mentioned, we use RMLMapper as a converter. The RML transformation rules for iXBRL reports, XBRL taxonomies, company profile, and officer profile can be found in this site.[29]

3.2.5.1 iXBRL report

To transform an iXBRL report into RDF, there are two sub-steps: 1) Converting by using RML rule and 2) Cleaning and constructing additional triples.

1) Converting by using RML rule

Based on the data model and ontology in the previous steps, we transform the financial report by using the RML mapping rule. Listing 1 C demonstrates fragments of the rule for iXBRL report in order to achieve the expected result (Listing 1 B).

– Listing 1 –

To begin this process, a Triples Map tag groups the mappings sharing the same subject (lines 1 and 18). Then we define the data source by specifying an `rml:logicalSource` element (lines 3 and 20). In each logical source, `rml:source`, `rml:iterator`, and `rml:referenceFormulation` are specified. Since the RML rule is per iXBRL instance, the input data source (Listing 1 A) and the query language – XPath are identical in both triple maps (lines 4, 6, 21, 23). The iterators are defined to iterate over `ix:nonNumeric` and `xbrli:period` tags (lines 5 and 22).

Next, we use the `rr:subjectMap` element to define the class of the subject. For example, to map entities in "factNonNumMapping" to the class `xbrll:Fact` and `iXBRLInstanceOnto: nonNumeric`, we use a blank node and the RML class property (lines 8 and 9). Afterwards, a set of predicate-object maps (`rr:predicateObjectMap`) elements define the creation of predicates and their object value for

[29] http://paul.ti.rw.fau.de/~un62efef/thesis/

the RDF subject. Lines 10 to 12 show the creation of an `xbrll:concept` predicate referencing the attribute `name` in `ix:nonNumeric` tag from iXBRL file (Listing 1 C). Lines 13 to 16 demonstrate the creation of triples when its object is the subject of another triples map, and the same source is used. Lastly, to integrate data from different sources/iterator, `rr:parentTriplesMap` is used as a reference through a join (`rr:joinCondition`) to another triples map (lines 13-16). In this case, we link the fact and the context by referring to the attribute `contextRef` value and attribute `id` value in the properties `rr:child` and `rr:parent`. The identical values are mapped so that the context – period can describe the fact.

For the second triple map (from line 18), we use a blank node as a subject. The predicates refer to a start date and end date of the fact reported (line 27-32). The proper data type – date is specified as shown in lines 29 and 32.

Once we have defined the mapping rules, an RML interpreter performs the data translation across the sources to RDF. The following step is to clean and standardize the data as well as to construct additional triples.

2) Cleaning and constructing additional triples

To provide information in a more user-friendly format, characters were removed, for example as \t, \r, and \n which are used to signify the end of a line of text and the start of a new one. In addition, appropriate data types were assigned to each value of the predicate `xbrll:value`, such as “20,000”^^`xsd:integer`, false, and “25 December 2020”^^`xsd:date`.

Prior to the entity linking of officer, a fact of the concept *NameEntityOfficer* is added with the predicate `rdfs:seeAlso` and the URI object `<http://paul.ti.rw.fau.de/~un62efef/ thesis/officer_profile.ttl#companynumber_officer_valuename>` as shown in Listing 2 A with the gray highlighted. This link is used to connect to the officer entity in the officer profile RDF file.

– Listing 2 –

Even though we adopt the lightweight vocabulary, we slightly change the structure and added existing vocabularies and our own created vocabularies in iXBRL instance

ontology to simplify the dataset e.g., removing blank nodes and indirect links. Examples of the vocabularies are `schema:startDate`, `schema:endDate`, `iXBRLInstanceOnto:hasInstant`, `iXBRLInstan ceOnto:hasDimensionAxis`, and `iXBRLInstanceOnto:hasDimensionValue`.

In addition, for the numeric items, there are iXBRL attributes such as `scale` to express the presented figure in the report. As shown in Listing 2 B, the value of *1452392* is applied with the scale of *zero* (i.e., unchanged value). Therefore, such value is translated to *1452392* of the predicate `iXBRLInstanceOnto:convertedValue`. In case the scale is not zero, we used a Python script to transform the value (i.e., multiple the value and the scale with the power of ten).

The fact with a value of dates was also transformed into the same format i.e., *25 December 2019, December 25, 2019*, and *25.12.2019* to *2019-12-25*.

3.2.5.2 XBRL taxonomy

In this study, we have consider the following four files for our transformation: 1) Schema file (XSD), 2) Label Linkbase (XML), 3) Reference Linkbase (XML), and 4) Presentation Linkbase (XML). These files were selected since the *Schema file* can provide us with the information about a concept itself, and the *Label Linkbase* provides a human readable label as well as a guideline description. While the *Reference Linkbase* refers to a concept to external resources, the *Presentation Linkbase* provides connections between concepts.

– Listing 3 –

The *Definition Linkbase* is not used as it also describes connections between concepts e.g., hypercubes and dimensions which allow companies to report more detailed information. For example, the concept *Rental leasing income* presents as the parent of 1) the concept *Income From Leasing Plant Equipment* and 2) the concept *Rental Income From Investment Property* in Presentation Linkbase.

On the other hand, the domain member is applied for those relationships in Definition Linkbase. The relationship of the first concept (1) is under the hypercube *Income main* and the relationship of the second concept (2) is under the hypercube *Income investment property*.

In the study of (Declereck & Krieger, 2006), the Definition Linkbase is used, however, the part-of/has-part properties are characterized through `xlink:from` and `xlink:to`. We decide to omit the Definition Linkbase to avoid any duplicate relationships and creation of additional triples. This is because our Presentation Linkbase can also express parent-child relations that are more or less comparable and our intention for this ontology is to be lightweight and easy to use.

Converting by using RML rule

Listing 3 C, demonstrates fragments of the rule for iXBRL report in order to achieve the expected result (Listing 3 A).

In this example, only the Schema file (XSD) and Reference Linkbase are demonstrated. For the complete transformation rule, please see this site[30].

The namespaces of elements and attributes in the Schema file should be replaced i.e., the input file (line 4). Similar to the previous mapping, the query language – XPath is specified (lines 6, 28, 41, 52, 64).

By specifying the expected result in RDF format, we use the `rr:subjectMap` element with a template-valued term map to define our subject to the class `skos:Concept` to map a URI of the concept. As the attribute `name` only contains a concept name, the taxonomy namespace is added in the URI to denote which taxonomy version referred to the concept (line 8 – before hash sign). The sets of predicate-object maps (`rr:predicateObjectMap`) elements (from lines 11) are used to create predicates and their object values for the RDF subject, namely `TaxonomyOnto:hasType`, `TaxonomyOnto:hasPeriod`, and `TaxonomyOnto:hasBalance` predicates. A term type is used to denote the IRI element.

From the second mapping (line 23), the Reference Linkbase file is used as an input source. The Reference Linkbase is a collection of referenceArc elements.

A referenceArc expresses a relationship between concepts and reference resources via the attributes `xlink:from` and `xlink:to`. The value of the attributes `xlink:from` and `xlink:to` must be equal to the value of the attribute `xlink:label`. If the value of the attributes `xlink:to` and `xlink:label`

[30] http://paul.ti.rw.fau.de/~un62efef/thesis/

matches, the predicate `dcterms:references` links from the attribute `xlink:from` to the corresponding external source (lines 42 and 56).

As some concepts may link to more than one external resources, the combination of the attributes `xlink:label` and `xlink:role` is used to uniquely link to different sources (lines 57 and 66).

Combining two taxonomies in different years

On completion of the transformation, the cleaning and restructuring steps of the triple are performed. Listing 4 shows an identical concept of 2014 FRC taxonomy (Listing 4 A) and in 2019 FRC taxonomy (Listing 4 B) in Turtle syntax.
As we can see, there are differences between the two taxonomies of the same concept, such as concept label and comment.

– Listing 4 –

To deal with 2014 and 2019 FRC taxonomies, we proposed to utilize an inferencing capability of semantic web that is the same-as relation (Listing 5). In the case that there are exact similar elements between such concepts in both years, this relationship can be defined with `owl:sameAs`. Otherwise, `rdfs:seeAlso` is applied to add additional information of the concepts in different years and avoid the inference.
It is required to normalize the URI concept by removing the taxonomy year, for example, `<http://example.org/core#TaxTaxCreditOnProfitOrLossOnOrdinaryActivities>`. In this way, both concepts are compared to each other whether there are any similar/different elements. After the comparison, the elements are categorized into *three different parts* as follows.

– Listing 5 –

1) *Unique 2014 concepts* (Listing 5 A) contain the elements of the concept which are specified in FRC taxonomy in 2014 only. The additional information can be found via `rdfs:seeAlso` with the normalized URI.
2) *Unique 2019 concepts* (Listing 5 B) contain the elements of the concept which are specified in FRC taxonomy in 2019 only. The `rdfs:seeAlso` with the normalized URI is also added in this file.
3) *Shared concepts* (Listing 5 A) contain identical elements between the concepts. In this case, the identical elements are the type and period.

In case that a concept in 2014 is removed in the 2019 taxonomy or a new concept is introduced in the 2019 taxonomy, there would be no shared concept URI created (i.e., `<http://example.org/...>`). Therefore, each concept stays in its own taxonomy year file.

A SPARQL query simplifies the query of information from a large number of companies in the dataset. By specifying a concept of any taxonomy year, the inferencing function can extract all the relevant information via the same-as relation.

3.2.6 Step 6: Link entities

After the transformation in Step 5, there should be three individual graphs of iXBRL reports (one graph per company filing), company profiles, and officer profiles. The XBRL taxonomy is considered as the ontology and therefore not included in this integration.

To link entities between graphs, a key identifier is required. Since every company has its own unique company number, the company number is used as a key identifier to match two sources. However, this information is not disclosed in the knowledge base like DBpedia. Alternatively, the company name can be used to perform the matching. While a company is identified by a number, there are no officer numbers to uniquely identify each officer. Therefore, a person name may be used as a key identifier to link a person entity. In addition, the information of a person's birth year should also be considered to facilitate the matching process. Table 3 shows the entity mappings of a company and an officer that we expected to achieve. The details are further described in each section.

Entity	Entity URI	Property	Linked entity URI
Company	http://paul.ti.rw.fau.de/~un62efef/thesis/iXBRL-File_NI004925.ttl#**entity_05572298**	owl:sameAs	company_profile:05572298
		owl:sameAs	http://business.data.gov.uk/id/company/05572298
		owl:sameAs	http://opencorporates.com/companies/gb/05572298
		owl:sameAs/ rdfs:seeAlso	wd:Q7245400
		rdfs:seeAlso	http://dbpedia.org/resource/Cobham_Hall
Officer	http://paul.ti.rw.fau.de/~un62efef/thesis/officer_profile#03213073_**officer_DJMorgan**	owl:sameAs	http://paul.ti.rw.fau.de/~un62efef/thesis/officer_profile.ttl#MORGAN_David_John
		owl:sameAs	http://dbpedia.org/resource/Samuel_Okikiol
		owl:sameAs	wd:Q7245400

Table 3: Example mappings of company and officer using owl:sameAs and rdfs:seeAlso

To connect two entities, we use `owl:sameAs` and `rdfs:seeAlso` relations. While `owl:sameAs` relation can be used to state that two individuals are identical (OWL Web Ontology Language Guide, 2004; OWL Web Ontology Language Overview, 2019), `rdfs:seeAlso` relation is used to indicate a resource that might provide additional information about the subject resource (RDF Schema 1.1., 2014). Therefore, the `owl:sameAs` relation is used if our matching can provide proof that they are the same entities e.g., via a company number or the Wikidata property of CH company ID. Otherwise, we apply `rdfs:seeAlso` to increase user confidence in graph usage and the quality of our graph.

3.2.6.1 Mapping company number

This approach is straightforward since a company number uniquely identifies the company. As mentioned in **Step 1**: Define goal and data requirement of the graph, both CH Linked Data service and Open Corporates provide company data in RDF format and URIs as identifiers of the company. Therefore, to connect our company entity in the iXBRL report with the company entity in those sources, we can simply use `owl:sameAs` with the URIs and corresponding company number generated.

In addition, in Wikidata, there is a property of CH Company ID (P2622) to denote the company number.

Therefore, we can also use `owl:sameAs` on the company entity with this ID property. For example, the company entity *00705392* from the financial report is linked to the company entity of CH, Open Corporates, and Wikidata via the same-as relation as shown below.

```
<http://paul.ti.rw.fau.de/~un62efef/thesis/iXBRL-
File_00705392.ttl#entity_00705392>   owl:sameAs
          <http://business.data.gov.uk/id/com-
     pany/00705392>,
          <http://opencorporates.com/compa-
          nies/gb/00705392>,
          <http://www.wikidata.org/entity/Q6798141>.
```

3.2.6.2 String similarity comparison

A strong string similarity is one of the methods used to perform entity matching (Shen et al., 2015). There are several techniques used in matching string data such as *character-based similarity metrics, token-based similarity metrics, phonetic similarity metrics, and numeric similarity metrics* (Wong, 2020).

While character-based similarity metrics (e.g., edit distance, affine gap distance, and Jaro distance metric) try to tackle typographical errors, rearrangement of words can be solved by using token-based similarity metrics (e.g., atomic strings) (Elmagarmid et al., 2007). Based on the dataset characteristics (e.g., length and structure), each algorithm has its strength and weakness in performing. Even methods that have been tuned and tested on a large number of previous matching problems can perform poorly on new and different matching problems (Bilenko et al., 2003, p. 20).

In particular, edit distance metrics or Levenshtein distance was commonly used – not only for text processing but also for biological sequence alignment (Bilenko et al., 2003). The character-based similarity metric like the Jaro-Winkler distance algorithm was also used in the study of (Lee et al., 2016) to calculate the string similarity score of a company name.

Based on our datasets, we applied three different approaches as illustrated in Figure 8. Two approaches were proposed for the officer entity linking, called prefix name mapping and birth year mapping.

For the company entity linking, the company names from two different sources were compared.

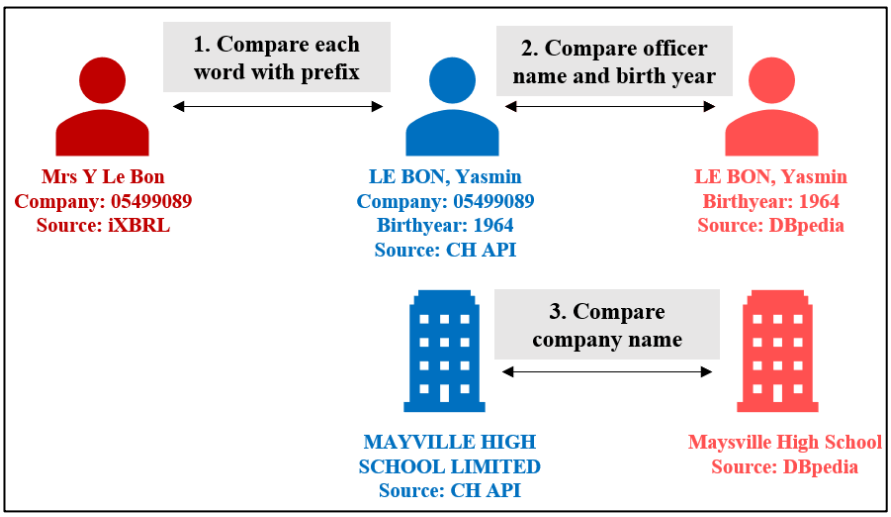


Figure 8: Summary of mapping methods

1. Officer – prefix name mapping

Prior to the comparison, the preprocessing task of the officer's name is required. By examining the officer name, the pattern of officer name in the iXBRL reports starts with a title, followed by abbreviated first and middle names, and last name, for example, *Mrs Y Le Bon*. Some do not start with the title but contain punctuation marks or symbols, such as *Graham Blackwell & Michael Budack*. The names reported in CH start with the last name in capital letters, followed by a comma and a first name, such as *CHARSLEY, Doris*.

After reviewing the officer's name, we perform the common preprocessing approaches to remove titles, education degrees, positions, and punctuations as summarized in Table 4.

Preprocessing methods	Name before preprocessing	Name after pre-processing	Target name
Remove prefix: mr, mr., miss, mrs, mrs., ms, ms., director, dr, dr., prof	Mrs Y Le Bon	Y Le Bon	LE BON, Yasmin
Remove suffix: bsc msc, bsc, msc, ma ba dip soc admin, bsc(hons) fca cta, ba (hons) fca, bsc mrics - vice chair, fca - vice chair, obe, ma, ca, dhc (chair-man), ma fca - chair, vice chair, (chairman), chair	J Douglas Craig OBE, MA, CA, DHC (Chairman)	J Douglas Craig	CRAIG, James Douglas
Remove/replace punctuation marks e.g., dash (-), point (.) and &	Graham Black-well & Michael Budack	Graham Black-well and Michael Budack	BLACKWELL, Graham
Separate the name after capi-tal letter	CE Haggerty, Da-vidChennells	C E Haggerty, David Chennells	HAGGERTY, Claire Elizabeth, CHENNELLS, David

Table 4: Summary of preprocessing methods for the entity linking

In addition, we noticed that there were cases where no space is in between the first and middle abbreviated names e.g., *CE Haggerty*. We, therefore, add space in between such name e.g., *C E Haggerty*. Even though there were some cases that the name was reported with all capital letters e.g., *TERENCE CASEY*, and adding a space between the capital letter could impact the matching result, it proofed to be a minority in the dataset.

For the first technique, **TextDistance**[31] **– Prefix** similarity is applied. The package verifies the longest match of the prefix, which suits our dataset's characteristics. The implementation consisted of comparing the initial name and accumulating the score of all comparisons. We select the name with the highest sum score as our best matching candidate.

To implement this technique, all letters are set to lower case and the name is split after space, i.e., *from Y Le Bon to y, le, bon*. The example of the implementation is shown in Listing 6.

[31] https://pypi.org/project/textdistance/

Referring to Listing 6 A there are three officers working in the same company as officer Y Le Bon. These three names are used to make the comparison. The prefix similarity is used to compare each individual name part (i.e., y, le, bon) against the split name of all three officers. In each comparison, a similarity score is generated. The higher the score, the more likely it is that the names are identical.

– Listing 6 –

As we can see in Listing 6 B, *y* is compared *three times* against each individual name part where the similarity score is *0.166* when comparing against *yasmin*. For each name, the scores of each comparison are accumulated. The name with the highest similarity score is the name to be mapped. In this case, the name *le con yasmin* is selected with the highest score at 2.166 while the score of the name *lee tannaz* is 0.667 and the score of the name *emma cook* is 0.

2. Officer – birth year mapping

To narrow down the amount of the comparison, we first check the birth year of both persons. By taking the data inconsistency between different sources into account, the birth year condition is relaxed to minus and plus one year.

For example, officer born in 1990 is compared to the candidate who was born between 1989 and 1991. To some extent, this approach is similar to Gawriljuk et al. (2016) study that checked candidates and target entities with birth years.

Based on the names in our dataset and DBpedia, the FuzzyWuzzy package is implemented. The punctuation marks removal is included in the package's feature. Since we prioritize accuracy above the number of matched results, the score threshold was set to 100. In total, there were 10 officers linked to DBpedia.

3. Company – name similarity comparison

In a similar way to the comparison between the officer's name from CH API and DBpedia, the company names between CH API and DBpedia were compared.

Before the comparison, the prefix and suffix of the company name were removed as summarized below.

The list of these prefixes and suffixes is mainly derived from the datasets.

- **Prefix:** the
- **Suffix:** company limited(the), limited(the), public limited company, companylimited , trust(the), limited.(the), (uk) co., ltd., (uk) ltd, (uk) limited, uk limited, company limited, co. limited, co ltd, co., ltd., company, limited, ltd, corporation, (incorporated), inc., (inc.), (company), p l c, plc, ltd., ltd., ltd.(the), llp, cic, c.i.c.

By removing prefixes and suffixes, the company names between the sources can be as similar as possible, and hence we can achieve more accurate results. The minimum threshold of comparison results was set at 100. In total, there were 438 distinct companies linked to DBpedia.

4 Evaluation

4.1 Graph quality

As explained in Step 6, the entity linking tasks searched and created links from the entities in the financial reports to the entities in the CH graph and external graphs – DBpedia and Wikidata. The similarity metrics were string-matching metrics (Levenshtein distance), comparing the names of companies and officers.

To assess the quality of the links, evaluation measures such as precision, recall, F1-measure, and accuracy are used as the assessment of entity linking (Shen et al., 2015; Burdick et al., 2011). For example, precision and recall are used in a scalable approach to incrementally building knowledge graphs (Gawriljuk et al., 2016) to measure the quality of the linking results in a knowledge graph of artists. To measure our graph, we, therefore, adopt precision, recall, and F1-measure. The calculation of each metric is described below (Shen et al., 2015).

- The *precision* determines the correctly linked entities by considering *all entity mentions that are linked* by the system. It is computed as the fraction of correctly linked entity mentions that are generated by the system. The higher the precision, the better the system is at ensuring that what is identified is correct.

$$Precision = \frac{correctly\ linked\ entity\ mentions}{linked\ mentions\ generated\ by\ system}$$

- The *recall* determines how correct linked entity mentions are by considering *all entity mentions that should be linked*.

It is computed as the fraction of correctly linked entity mentions to total entity mentions that should be linked. This metric measures how many of the items that should have been identified were identified.

$$Recall = \frac{correctly\ linked\ entity\ mentions}{entity\ mentions\ that\ should\ be\ linked}$$

In this case, the number of entity mentions that should be linked is equal to the total entities. As the accuracy is calculated by using the number of correctly linked entity mentions divided by the total number of entity mentions, the recall ratio is equal to the accuracy.

- In *F1-measure*, precision and recall measures are combined to produce a single measurement. F1-measure is defined as the harmonic mean of precision and recall.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Based on the calculation, the evaluation results of the entity linking for officers are summarized in Table 5. The number of correct mappings was manually checked by the author. For example, we conclude that E R R Martin and Martin Edward Robert Randall are identical and mapped. In case that the name with the maximum score was ambiguous, further information on the CH website is verified and conclude whether it is correct such as G Baker and Brock William George.

Mapping methods	Mapping sources	Total entities	Entities found	Correct mapping	Precision	Recall	F1
Officer – prefix name mapping	iXBRL □ CH	1.704 □ 4.331	1.700	1.675	98,53%	98,30%	98,41%
Officer – name similarity comparison	CH □ DBpedia	4.331 □ 743.584	10	9	90,00%	0,21%	0,41%

Table 5: Evaluation results of the entity linking for officers

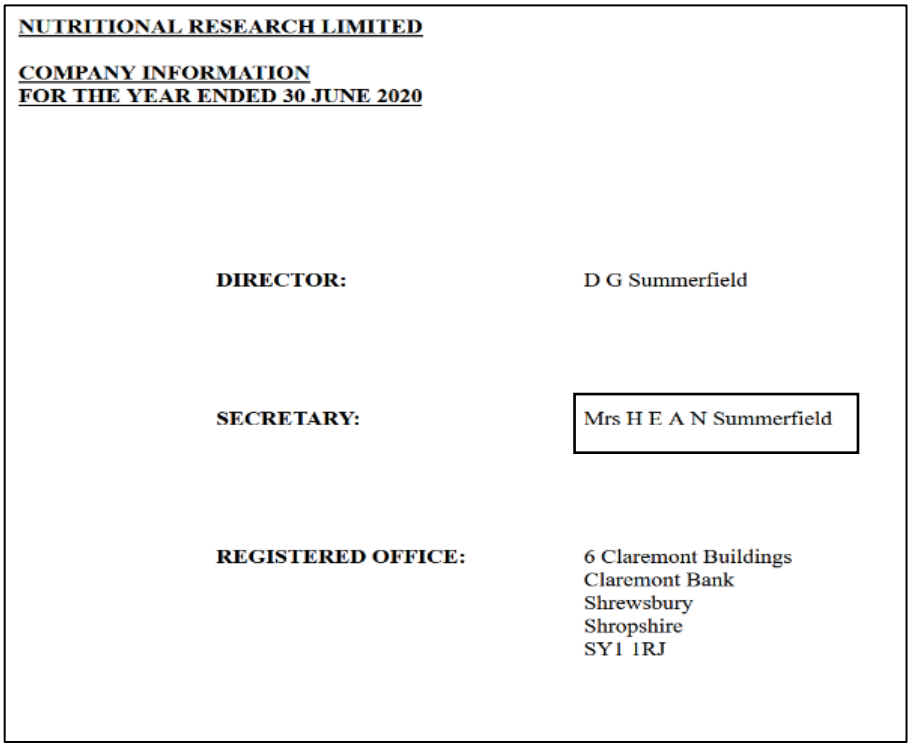
The first approach is where we map the officers between the converted financial data (1.704 officers) and CH (4.331 officers) by using the officer prefix names. In total, there are 1,700 entities found with the correct mapping of 1.675 entities.

Four entities are not matched since the string comparison score was zero and there were no such officers in the CH website. All evaluation metrics report high percentages of about 98 percent. Due to the strict high threshold set, a high precision score was achieved. After checking the results, the false candidates of 25 officer entities can be categorized as follows.

1) *False candidate:* 19 officers names were matched to different officers.
2) *Missing candidate:* 6 officers names were not found on the CH website.

In the false candidate case, there is one case that the names of two officers were combined. For example, the *director D G Summerfiel* and the *secretary Mrs H E A N Summerfield* are reported in the financial report (Figure 9 A). However, our algorithm is unable to match these properly since the secretary's name shown on the CH website is *NOBLE, Helen Elizabeth Ann*, which does not include *Summerfield* (Figure 9 B).

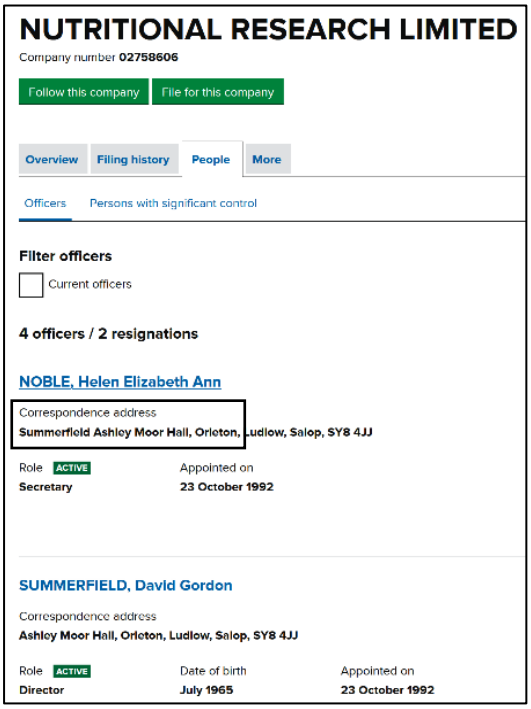
Figure 9: Mismatched officer's name between iXBRL report and CH website


NUTRITIONAL RESEARCH LIMITED

COMPANY INFORMATION
FOR THE YEAR ENDED 30 JUNE 2020

DIRECTOR:	D G Summerfield
SECRETARY:	Mrs H E A N Summerfield
REGISTERED OFFICE:	6 Claremont Buildings Claremont Bank Shrewsbury Shropshire SY1 1RJ

a) Information from iXBRL report



b) Information from CH website

For the second approach, the attempt to connect officer entities from CH (4.331 officers) to DBpedia (743.584 person entities) shows high precision of 90 percent i.e., one false candidate, but poor results of recall and F-1 metrics, 0.21 and 0.41 percent, respectively.

The most likely causes of a low number of entities found are the high threshold set for the similarity score and our small dataset. In addition, due to the business size of reporting entities (i.e., SME), it is likely that the officer entities who work in the companies have not been created in DBpedia yet.

For the company name comparison, we perform the string matching between the company names obtained from CH (2.846 companies) and DBpedia (421.432 company entities). As we can see from Table 6, about 15 percent (i.e., 438 entities from 2.846 entities) were identified. To enhance the confidence of this raph, the strict match at 100 percent was applied which may cause the low matching. As mentioned in Section 3.2.6.2 (Company – name similarity comparison), we used `rdfs:seeAlso` to avoid the incorrect inferencing. Therefore, the quality metrics are not applied in this case.

Mapping methods	Mapping sources	Total entities	Entities found	Linked entities (`rdfs:seeAlso`)
Company – name similarity comparison	CH → DBpedia	2.846 → 421.432	438	15,39%

Table 6: Evaluation results of the entity linking for companies

In addition, to evaluate whether our results can provide confidence to the users, we use OpenRefine[32] to link the officers and companies with Wikidata entities. OpenRefine is a tool to work with messy data such as data exploration, clean and transform data, and reconcile and match data. Reconciling feature is to match a dataset with the dataset from an external source such as Wikidata or a local dataset. It is a semi-automated function where it matches a cell value at its best but still requires human judgment. Therefore, after the matching, the results are to be reviewed and approved (Reconciling, 2021).

For the reconciliation with Wikidata, the user must define the entity type, the resource class, and optionally attributes that will provide higher precision at the discovered links (Wikidata: Tools/OpenRefine, 2021). We can choose a type to reconcile, for example, Human (Q5) or Organization (Q43229). This type and its subtypes will be included in the reconciliation. Therefore, if the Organization type is selected, OpenRefine also matches items against Business (Q4830453), Public school (Q2015541), Limited liability partnership (Q1588658), etc. which are the subclasses of the Organization.

For the linking process between companies from our RDF dataset of CH and Wikidata, the results are 346 links for the whole dataset. On the other hand, the results from the linking process between officers are 25 links for the whole dataset.

While the results of the companies are likely to be correctly matching, we are not confident with the person matching. This is due to the fact that there is no sufficient information to confirm whether they were identical persons.

[32] https://openrefine.org/

Based on the study of (Angelis & Kotis, 2021) which utilized the external tools to perform the linking such as OpenRefine and Silk, there were lacks of performance and choices during the linking process. Therefore, the front-end web application is developed in the study where it is designed to perform the linking tasks in a customized way.

4.2 Data analysis

The following parts present the questions and results from the SPARQL queries. There are various approaches to express a query such as using the `SELECT` or `CONSTRUCT` form. In our work, we demonstrate different types of queries which are included in the specified Appendix.

4.2.1 Financial analysis

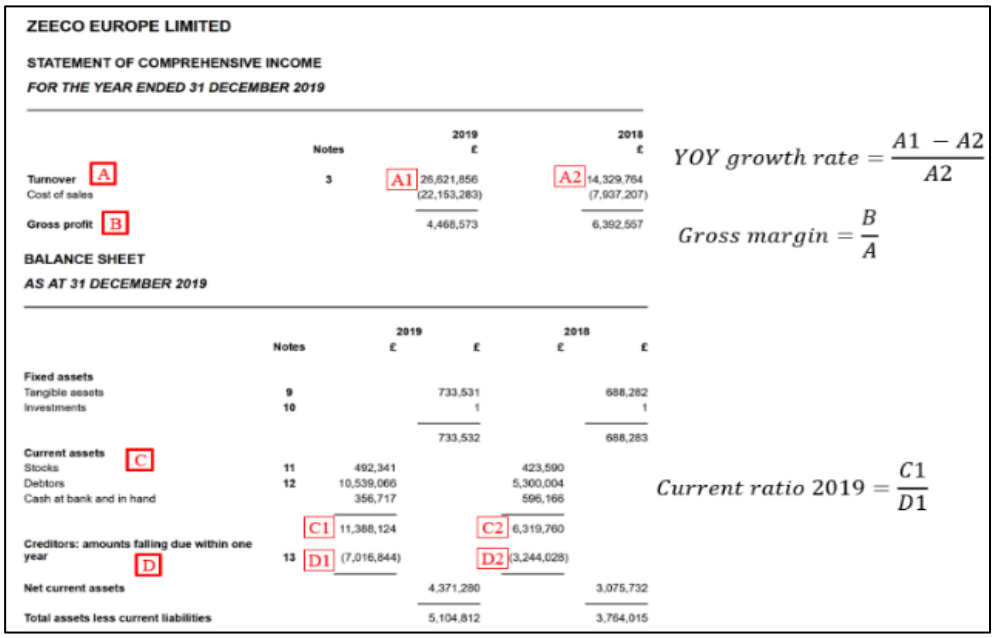
Financial analysis based on figures in financial statements can indicate if the business's financial future is secured or not. An evaluation of the financial performance and condition of a business concern can be identified via financial ratios. The measurement of profitability and liquidity is, therefore, an essential component of the financial health of a company (Chukwunweike, 2014).

For example, current profitability can provide information about the potential of the company to internally generate funds through operating activities. Two-year data may suggest either a positive or negative signal (Piotroski, 2000).

As a result, the questions listed below are used as our competency questions for this financial analysis. These questions were expressed in the queries from both balance sheet and profit and loss statements.

- What is the revenue of companies in 2018 and 2019 and the YOY growth?
- What is the gross margin ratio of companies in 2018 and 2019?
- What is the current ratio of companies in 2018 and 2019?

Figure 10: Example of SPARQL query of financial ratio



ZEECO EUROPE LIMITED

STATEMENT OF COMPREHENSIVE INCOME

FOR THE YEAR ENDED 31 DECEMBER 2019

	Notes	2019 £	2018 £
Turnover A	3	A1 26,621,856	A2 14,329,764
Cost of sales		(22,153,283)	(7,937,207)
Gross profit B		4,468,573	6,392,557

$$YOY\ growth\ rate = \frac{A1 - A2}{A2}$$

$$Gross\ margin = \frac{B}{A}$$

BALANCE SHEET

AS AT 31 DECEMBER 2019

	Notes	2019 £	£	2018 £	£
Fixed assets					
Tangible assets	9		733,531		688,282
Investments	10		1		1
			733,532		688,283
Current assets C					
Stocks	11	492,341		423,590	
Debtors	12	10,539,066		5,300,004	
Cash at bank and in hand		356,717		596,166	
		C1 11,388,124		C2 6,319,760	
Creditors: amounts falling due within one year D	13	D1 (7,016,844)		D2 (3,244,028)	
Net current assets			4,371,280		3,075,732
Total assets less current liabilities			5,104,812		3,764,015

$$Current\ ratio\ 2019 = \frac{C1}{D1}$$

a) A financial report with ratio calculation

```
CONSTRUCT {
    ?comp ex:hasRevenueGrowthrateY19_18 ?growth_R }
WHERE {
 ?comp  iXBRLInstanceOnto:hasFact [xbrll:concept ?turnover;
                                   schema:endDate ?endDate_prev;
                          iXBRLInstanceOnto:convertedValue ?valprev_R].

 ?comp  iXBRLInstanceOnto:hasFact  [xbrll:concept ?turnover;
                                    schema:endDate ?endDate;
                            iXBRLInstanceOnto:convertedValue ?val_R].

 FILTER ((?endDate_prev > "2018-06-30"^^xsd:date && ?endDate_prev <=
"2019-06-30"^^xsd:date) && (?endDate > "2019-06-30"^^xsd:date &&
?endDate <= "2020-06-30"^^xsd:date))

 VALUES ?turnover {<http://xbrl.frc.org.uk/fr/2014-09-
01/core#TurnoverRevenue>
<http://xbrl.frc.org.uk/fr/2019-01-01/core#TurnoverRevenue>
<core:TurnoverRevenue> }

BIND ((((?val_R - ?valprev_R)/ ?valprev_R)*100000)/1000 AS
        ?growth_R)
FILTER ( ?valprev_R > 0)
```

b) A SPARQL query of YOY revenue growth

The calculation of gross margin ratio uses gross margin scaled by total sales (Piotroski, 2000), while the company's current assets are compared to its current liabilities at fiscal

year-end for the current ratio or liquidity ratio (Piotroski, 2000). In this example (Figure 10), a financial statement, as well as the calculation formula, are shown in Figure 10 A. We choose the `CONSTRUCT` query (Figure 10 B) to express the question on YOY revenue growth during the financial periods of 2018 and 2019. The periods can be specified in the `FILTER` which the ended periods after 30th June 2018 and before 30th June 2019 denote the financial period of 2018. In the same way, the same dates of the later year are applied for the 2019 period. The concept *TurnoverRevenue* of taxonomies from different years is placed in `VALUES`. The calculation is performed by using `BIND` which the divider value should be more than zero.

– Listing 7 –

Listing 7 shows the query result.[33] From the result, an example of two companies in our database which have the information of revenue, gross margin, current assets, and current liabilities. From the figures, we can see that the revenue growth slightly decreased at -3.17 percent for the first company. Contrary, the significant positive growth trend in revenue at almost 86 percent and high gross margin for the second company. An improvement in margins indicates a potential improvement in factor costs, a reduction in inventory costs, or a rise in the price of the firm's product (Piotroski, 2000).

An improvement of the current ratio or liquidity is a good signal about the firm's ability to service current debt obligation (Piotroski, 2000). As shown in the listing above, the current ratio in 2018 of the first company and the current ratio in 2018 and 2019 of the second company are more than 1.0. It indicates that the companies had a larger proportion of short-term asset value relative to the value of their short-term liabilities. However, the first company may appear to be struggling to pay its bills in 2019 due to the lower ratio below 1.0.

Moreover, further analysis could be performed regarding the correlation between profitability and current ratio. As mentioned in Chukwunweike's (2014) study, a positive correlation is from idle funds borrowed by a business generate profit and lower business costs.

[33] refer to the SPARQL query in Appendix B – SPARQL Query for Financial Analysis.

In addition, results from the queries can lead to further analysis. For example, significant growth or unusual profitability, operating losses, related party transactions beyond ordinary, significant operations across international borders are examples of fraud risk factors (Skousen et al., 2009).

4.2.2 Industry analysis

An industry analysis examines how the business in the market is competitive in terms of sizes, earnings, and capital and human resources. Referring to Table 7, the first question aims to show how to group the companies with the number of accounting items, which in this case we chose the company's asset with the amount of GBP 10.000 in 2019 for the partition. The second question is where we would like to know the average profit of companies registered in a specific industry code. The last question provides the top five industries with the largest average number of employees.

In this example, we chose a `SELECT` query to express the questions.[34] The query results are presented in Table 7.

[34] refer to the SPARQL query in Appendix C – SPARQL query for industry analysis.

Competency questions	Queried results
How many companies have an asset less than and more than GBP 10.000 in 2019?	"assets < 10000" "82"^^xsd:integer "assets >= 10000" "111"^^xsd:integer
What is the average profit of companies registered in SIC code 68209[35]?	"62799.7"^^xsd:decimal
What are the top 5 industries (SIC code) which have the largest average number of employees in 2019?	1. sic:59111 "Motion picture production activities"@en " 505.75"^^xsd:decimal 2. sic:96090 "Other personal service activities n.e.c."@en " 287.52"^^xsd:decimal 3. sic:47760 "Retail sale of flowers, plants, seeds, fertilisers, pet animals and pet food in specialised stores"@en "102.0"^^xsd:decimal 4. sic:28290 " Manufacture of other general-purpose machinery n.e.c."@en "80.0"^^xsd:decimal 5. sic:78200 "Temporary employment agency activities"@en "78.0"^^xsd:decimal

Table 7: Competency questions and queried results of industry analysis

As can be seen from the first question in Table 7, although our dataset comprises more than one thousand reports, not all companies disclose all accounting items or submit full reports to CH.

Based on information from the UK government, small companies (i.e., with 0 to 49 employees) can send abridged accounts to CH, and micro-entities (i.e., less than 10 employees) can prepare simpler accounts only to meet the minimum requirements and send only balance sheet with less information to CH (Prepare annual accounts for a private limited company, 2015). In 2020, these companies account for 99.3 percent of the business population in the UK (Business population estimates for the UK and the regions 2020, 2020).

The second question is where we are interested in the average profit of companies that are in the real estate business. According to the result, the average profit is around GBP 63.000. By knowing this information, we can also combine this information with further one from the UK government e.g., industry size and turnover.
For example, the SIC code in the query can be changed to other industry sectors. In

[35] Standard Industrial Classification code 68209 – Letting and operating of own or leased real estate.

addition, this SIC code information can be leveraged to find an interesting insight such as the company's competitors.

The last question is to observe the employment area. Based on the result, it is likely that the companies in the top two sectors, namely *Motion picture production activities*, and *Other personal service activities* are large size businesses (i.e., more than 250 employees) and the other three sectors are medium-sized enterprises (i.e., with 50 to 249 employees). From the information provided by the UK government, the largest SME employment in 2020 was in the Wholesale and Retail Trade and Repair sector. Which industry is on the rise can be presented through the growing number of employment, compared with the previous year's number.

4.2.3 Company analysis

In this analysis, we retrieve further information from financial reports. In addition, it can help clarify the ambiguous data as presented in Section 1.2 – Research Objective. As a result, the questions listed below are used as our competency questions for this company analysis.

- Which companies have parent and/or subsidiary companies?
- What is the full name of Mr. J D Millet – the director of the company Newgate Kennels Limited? Is there any further information that we can learn more about this person e.g., birthday, nationality, occupation, and address?

In this example, we chose a `CONSTRUCT` query to express the questions.[36] The query results are presented in Table 8.

[36] refer to the SPARQL query in Appendix D – SPARQL query for company analysis.

Competency questions	Subjects	Predicates	Objects
Which companies have parent and/or subsidiary companies?	**company_profile:BR007873**	schema:parentOrganization	<http://business.data.gov.uk/id/company/FC025590>
What is the full name of Mr. J D Millet? Is there any further information that can we know about this person e.g., birthday, nationality, occupation, and address?	**officer_profile:MILLETT__Joel_Duncan**	rdf:type	foaf:Person, org:Membership
		foaf:name	"MILLETT, Joel Duncan"
		dbo:birthYear	"1961-01-01"^^xsd:gYear
		OfficerProfile-Onto:hasBirthday_Month	1
		OfficerProfile-Onto:hasOfficerAppointment	<http://paul.ti.rw.fau.de/~un62efef/thesis/officer_profile.ttl/company/00634835/appointments/mkevTjwtnsfE0gSt3ya_TkQhRpQ>
	<http://paul.ti.rw.fau.de/~un62efef/thesis/officer_profile.ttl/company/00634835/appointments/mkevTjwtnsfE0gSt3ya_TkQhRpQ>	OfficerProfile-Onto:isAppointedOn	"1995-11-30"^^xsd:date
		schema:nationality	"British"
		schema:hasOccupation	"Kennels Manager"^^xsd:string
		locn:address	<http://paul.ti.rw.fau.de/~un62efef/thesis/officer_profile.ttl#address_/company/00634835/appointments/mkevTjwtnsfE0gSt3ya_TkQhRpQ>
	<http://paul.ti.rw.fau.de/~un62efef/thesis/officer_profile.ttl#address_/company/00634835/appointments/mkevTjwtnsfE0gSt3ya_TkQhRpQ>	locn:addressArea	"Wilmslow"
		locn:thoroughfare	"Newgate House Newgate Road"
		locn:postCode	"SK9 5LL"

Table 8: Queries and results for company analysis questions

Regarding the first question, the result shows only one relationship between the parent of company number *BR007873* and the entity <http://business.data.gov.uk/id/company/FC025590>. More relations could be expected for a larger dataset. The second question is to find additional information about the officer *JD Millet*, which is the name reported in the financial statements. From the integration task, the graph can provide additional information through the link specified using the same-as relation. In this case, not only personal information (e.g., birth year, occupation, nationality, and address) is revealed, but also the appointment information such as the appointed date and the resigned date, if any.

Overall, these results indicate that our graph was able to answer all competency questions and therefore achieved the second goal of analysis questions.

4.3 Data visualization

Turning to our final goal – data visualization, not only the graph can generate correct data to support the analysis questions, but it is also possible to visualize the generated data as a variety of graph types. In the semantic web, users can use data visualization to explore the content of the data, discover interesting patterns, infer correlations and causalities, and assist sense-making tasks (Bikakis & Sellis, 2016). It also allows non-domain and non-technical users to understand the structure. Access to the web of data can be achieved with a clear and coherent visualization of Linked Data, which can also increase its use outside the semantic web community (Dadzie & Rowe, 2011). In this way, the users can formulate queries, identify links between resources, and find new information. Simply stated, it has the potential to reduce the technical barrier and make the Web of Data accessible to everyone (Dadzie & Rowe, 2011).

Allowing people to interact with Linked Data is an important and challenging step to move the semantic web forward. It is still unclear how to effectively support people to query, browse and visualize the data (Santo & Holzer, 2020).

Nowadays, there is a range of tools to visualize Linked Data, including visualRDF,[37] rdfdot,[38] and ontology-visualization.[39] However, programming-related skills, as well as Linked Data knowledge, might be needed for installation and employment.

As (Dadzie and Rowe, 2011) reminds us of "1 Picture ≈ 1K words", one picture can be presented through a number of visualization techniques such as pie and bar charts, scatter plots, tree and graph visualizations, matrices, timeline, map and landscape views, bubble and sunburst plots, and iconography. Text-based visualization such as tag clouds and phase nets can support an analysis of the text. Well-known tools such as Tableau[40] and Protovis[41] are used in novel visualization applications.

To present users with the data generated from the graph, we have built interactive maps in a widely known *business intelligence software – Tableau*, and *mapping platform – Google Maps* in Section 4.3.1. In addition, a graph visualization is provided by using software called "Tarsier" in Section 4.3.2.

4.3.1 Map visualization

An interactive map can display our data as locations with additional insight. The map also allows people and machines to interact and supports the use of filter functions for a range of data analyses. The filtered results can be recalculated and shown in real-time. As noted above, our visualization tools are Tableau and Google Maps.

Tableau, founded in 2003, is the world's leading analytics platform with the supporting features to explore and manage data, and discover and share insights. The platform supports various types of data visualizations such as histograms, tree maps, box plots, and maps. To analyze data geographically, we can plot the data on a map in Tableau. In this way, it allows us to understand the trends or patterns in our data.
Types of maps offered in Tableau include proportional symbol maps, heatmaps, flow maps, and spider maps (Build maps in Tableau, n.d,).

[37] https://github.com/alangrafu/visualRDF.

[38] https://github.com/wastl/rdfdot.

[39] https://github.com/usc-isi-i2/ontology-visualization.

[40] https://www.tableau.com/.

[41] https://mbostock.github.io/protovis/.

To build a map in Tableau, location data such as location names, or latitude and longitude coordinates are required in the data source (Build a simple map, 2021). Tableau can also recognize the location data for building map views such as worldwide airport codes, cities, countries, regions, territories, states, provinces, and some postcodes (Location Data that Tableau Supports for Building Map Views, 2021).

Our graph permits a map visualization by linking the financial data and company data. To collect all the required data in our dataset, we can use the SPARQL query. Starting from our financial statements, by using the same-as relationship linked across multiple data sources, the latitude and longitude information can be retrieved through the CH SPARQL endpoint service. To provide useful insight to the users, additional information was also retrieved, for example, company name, company ID, number of employees, period of information, business activity(ies) including SIC code(s), and website links to company and officer information.

After retrieving the results in CSV format, it was necessary to check and clean the retrieved data. For example, from the triple store, the textual information about business activity includes commas. This might cause confusion in the CSV format, which uses commas as a separator, such as the business activity “Water collection, treatment and supply”. Therefore, the commas needed to be replaced with space or other appropriate punctuations. Once the data is cleaned, the file is imported into the software.

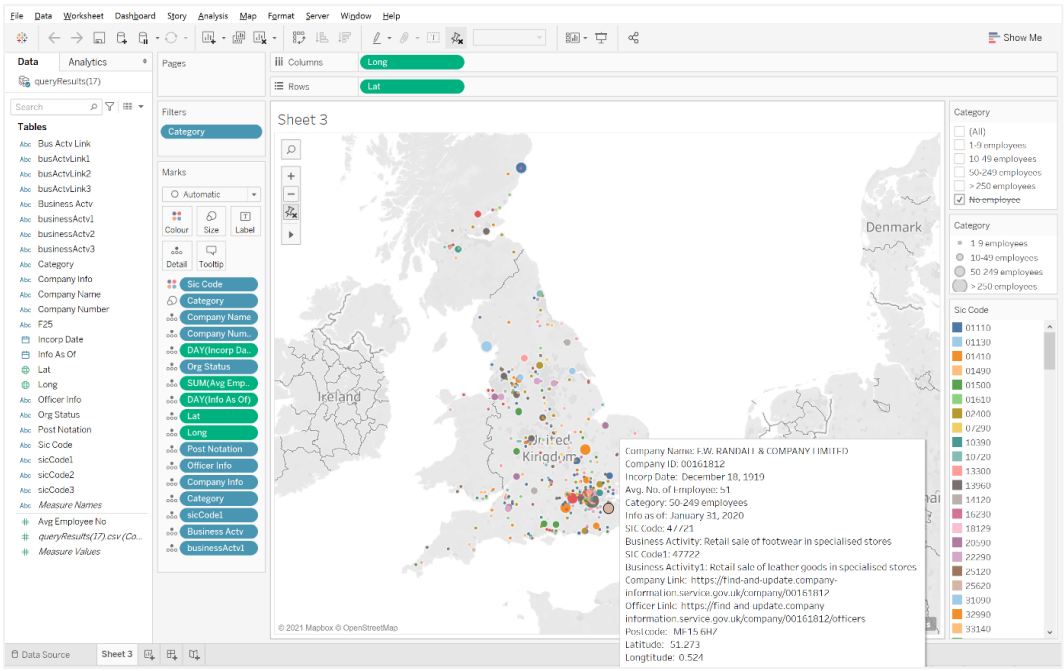


Figure 11: Interactive map visualization in Tableau

Figure 11 shows a wide distribution of the companies in our dataset across the UK. As we can see, the information of an average number of employees is divided into five groups (starting from zero employees to more than 250 employees), similar to the statistical data provided by the UK government. Therefore, we plot the data and size them by the categories. In addition, the color of circles varies from business activities – SIC codes.

In this example, we filtered out the companies which have zero employees, presented on the upper left corner of the map. The maps created in Tableau are accessible, interactive, and shareable via Tableau Online,[42] Tableau Public,[43] and Tableau Server.[44] For more information, please refer to our map.[45]

We can see that in this case, the businesses are mainly located in the south of the UK and London. Statistics also show that 35 percent of the businesses are located in those areas, specifically 1.1 million businesses in London and 0.93 million businesses in the South East of England (Business population estimates for the UK and the regions 2020, 2020). The color of SIC codes used to distinguish the industries reflect that the businesses carried out in the UK are diversified. The SIC code helps CH to see trends emerge and to track the health of different parts of the economy (Townley, 2018).

In addition, by selecting the circle, more information about a company can be revealed. In this case, the selected company *F.W. RANDALL & COMPANY LIMITED* was established around 100 years ago. The company engages in two business activities: retail sales of footwear and leather goods. As of January 2020, this company employs 51 employees and is classified as a medium-sized enterprise.

Another possibility to visualize a map is via Google Maps[46]. The features provided by Google Maps are such as planning trips, measuring distances, getting coordinates, searching nearby, and contributing to the map (Tips and tricks for Google Maps on your computer, 2021). We can create our map based on the dataset we have in varied formats e.g., CSV, TSV, XLSX, and Google Sheet. To properly visualize data in the

[42] https://www.tableau.com/products/cloud-bi.

[43] https://www.tableau.com/products/public.

[44] https://www.tableau.com/products/server.

[45] https://public.tableau.com/app/profile/pk5592/viz/Thesis_16226251152540/Sheet3.

[46] https://www.google.de/maps.

map, latitude-longitude information, addresses, or place names are required. In addition, the import file is restricted to contain only less than 2.000 rows (Import map features from a file, 2021).

As our data is in CSV format with the geographical information, we can simply import the file into the application. In contrast to Tableau, it is not possible to show the size of the data.

Therefore, in this task, we divided the data into different files based on the categories of employee numbers. After we imported those files, the icon and color can be adjusted based on the layer of the data on the map. The use of icons and colors are options available for overlaying multiple attributes on a single map. For example, those large companies (i.e., more than 250 people) are in purple and other business sizes are coded in different colors as shown in the layer box on the left of Figure 12. In this way, we can clearly see that the majority of businesses in the UK have 1 to 9 employees.

Moreover, each pin is embedded with the data we imported. In this case, the information of company *KCM Marketing (UK) Limited* is presented such as the company number, the number of employees, the incorporate date, and business activity(ies). For further information about the company or its officers, we can simply click on the gov.uk website provided.

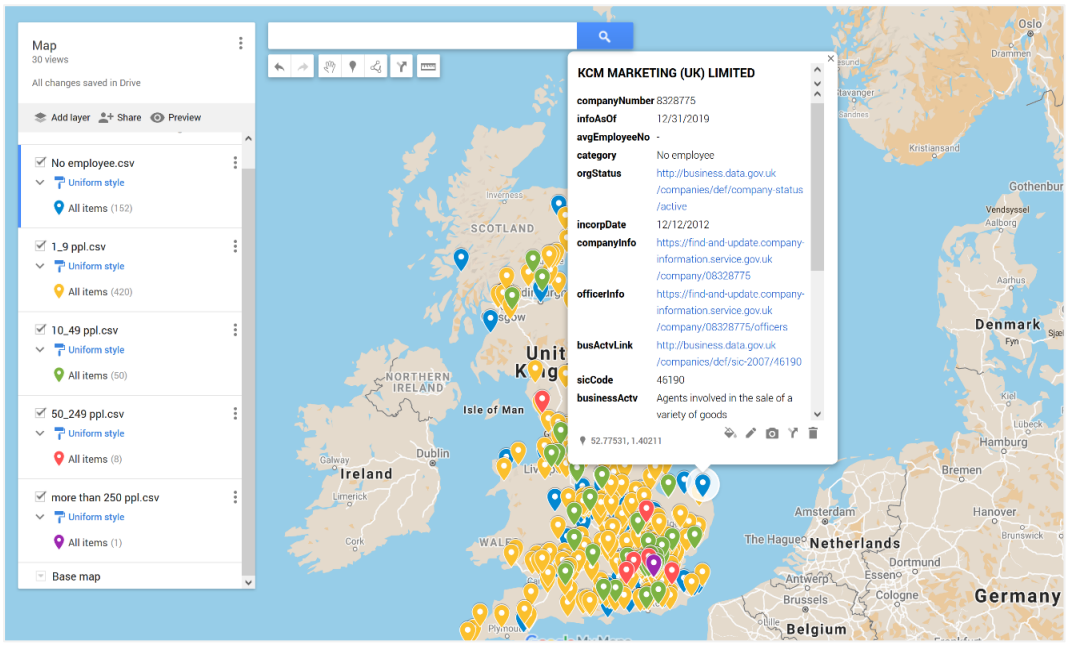


Figure 12: Interactive map visualization in Google Maps

Importantly, it is possible to share the map, search for the company, as well as find the direction. People who are interested can thus utilize and edit it. For more information, please refer to the map[47].

4.3.2 Graph visualization

Apart from using the business intelligence tool and map application, this part illustrates our linked data in a graph network. As described in Approaches to visualising Linked Data: A survey (Dadzie & Rowe, 2011), node-link tree and graph visualizations, in both 2D and 3D, are widely used to represent hierarchically structured data such as ontologies or networks. As previously mentioned, our graph visualization tool is Tarsier.

Tarsier[48] is a tool to interactively visualize RDF graphs in 3D. It adopts the concept of semantic planes to allow partitioning and highlighting links in order to evaluate data from a better viewpoint. The semantics planes or layers are designed to organize a set of resources sharing a set of common semantic features. The purposes of this tool are to assist inexperienced users in learning new data representation formats and to improve inspection capabilities so that relations may be discovered even in complex RDF graphs (Viola et al., 2018).

To use Tarsier, data from desired datasets are required to be connected to a SPARQL endpoint. In this paper, we deployed Blazegraph[49] as our SPARQL endpoint as we were unable to connect via Apache Jena Fuseki, which we use for all querying tasks.

Using Tarsier, we want to visually answer the questions: "*If the officers working in company number 01070684 also working in other companies?"*

From this query, it provided us the answer to our question in RDF format, for example, officer *Dollen Brian* is an employee in four companies. We can visualize this data in Tarsier as shown in Figure 13.

[47] https://www.google.com/maps/d/u/0/edit?mid=1DkH3cQbdpEIs1BW8qVDieGV6DWsTY4f&usp=sharing.

[48] https://github.com/desmovalvo/tarsier.

[49] https://blazegraph.com/.

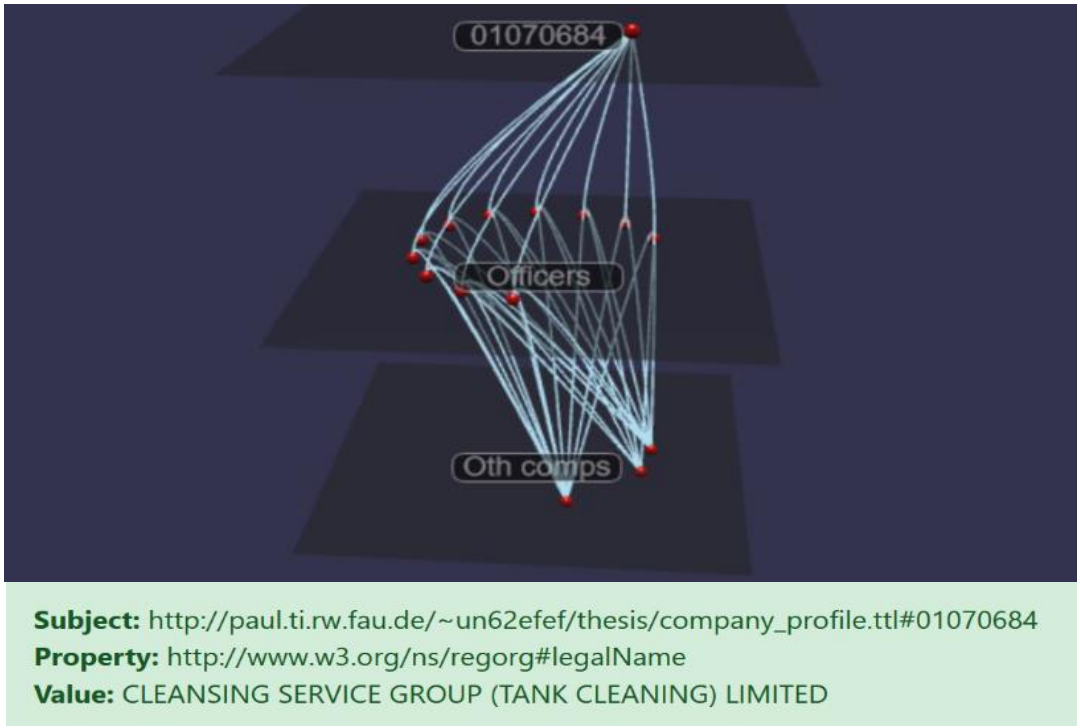


Figure 13: Semantic plane visualization in Tarsier

In this graph visualization, we created the highest semantics plane containing the company number *01070684* which we are interested in. When we click at the node, it provides us the company name *Cleansing Service Group (Tank Cleaning) Limited*, shown in the green box below the figure. The second plane is created containing the officers working in the company. This plane shows the relationship between the subject and the objects of a `schema:member` triple. This can be done by selecting the object property `schema:member` and clicking on Raise (S and O). From the figure, there are 11 employees who are members of this company. Finally, the third plane presents the companies in which those officers are the employees. Interestingly, they are mostly members of the other three companies. It could be that these officers have been appointed in different periods.

Further questions are such as if those companies are in the same group or what is the relationship among the officers. In addition, questions such as the company's group structure should be attempted to visualize using Tarsier.

Together, these visualization results support the notion that the combination between business and linked data.

5 Conclusion

The digital transformation involves almost all businesses and industry sectors including financial and accounting such as the implementation of financial filings in XBRL and iXBRL formats worldwide. Numerous benefits can be achieved from the growing number of companies implementing the filings and regulatory bodies developing the filings. There are studies related to linked data and the financial domain which proposed that RDF can be a common representation for data integration. Even though is a number of studies regarding XBRL with the semantic web technologies, the study of those with iXBRL is still limited. Therefore, the aim of the present study is to present how current linked data technologies can be used to support business people, in particular iXBRL financial reports. In this paper, there are three goals set to be achieved: graph quality, data analysis, and data visualization.

Starting with the background, Section 2 provides the extensive technical and non-technical background of both iXBRL and semantic web technologies. Previous studies regarding the transformation of the financial statements, the ontologies, and the entity linking are reviewed in Section 3. Our identified gaps we attempt to fill are the transformation of iXBRL into RDF using the RML tool, the alternative solution for taxonomies in different years, and for the entity linking.

In total, there are seven steps to create the knowledge graph. To build our graph and to develop the ontologies, we reuse various existing ontologies describing the company and officer information. This step-by-step instruction can be used to create other knowledge graphs with some modifications based on datasets and data sources. Since the latest taxonomy may be different from the previous version, we also propose the alternative in dealing with the taxonomies in different years.

This approach uses the inferencing capability of the semantic web to facilitate when performing SPARQL queries. Three different approaches of string comparison are proposed for officer and company entities. The last step is to perform the evaluation based on the paper goals.

Based on the first goal of graph quality, the results of this finding show that our graph is able to provide the correct information with the high precision and recall metrics compared to another method – using the external tool (OpenRefine). By carefully using the linking relations – `rdfs:seeAlso` and `owl:sameAs`, it can prevent users

from false assumptions or conclusions with regards to the inferencing resources attached to such entities. In addition, it allows users to discover more information from such links. The second goal is also achieved from the queries expressed to answer our analysis questions. Users can easily adjust the queries or express new queries based on their analysis questions. For the third goal, use cases involving map and graph visualizations are illustrated with the tools – Tableau, Google Maps, and Tarsier. The patterns of the data as well as the relationship between the data from different sources are captured through both presentations.

Taken together, these findings provide a potential mechanism for utilizing RML tools with financial data integration, handling the taxonomies in different years, and the entity linking techniques. As there are a growing number of countries, government bodies, and companies implementing reports in iXBRL format, we believe that our proposed approach can be a solution to this matter.

The most important limitation lies in the fact that the dataset obtained was relatively small. By increasing the number of reports, a different result of our graph quality should be expected, especially the precision and recall metrics when linking entities. The data quality from information disclosed in iXBRL and CH website also affected the evaluation result, as addressed in the result in Section 5.1. In addition, the small sample size did not allow the discovery of a variety of patterns or relationships in the data analysis.

Despite the limited dataset, one limitation we encounter is during the attempt to apply semantic reasoning on the integrated data in order to infer new relationships between the data.

The inferencing process with the integrated reasoning services of Apache Jena Fuseki can not be completed for our RDF graphs possibly due to a lack of our computational capacity. As a result, the complete taxonomies in RDF with the same-as relation could not be examined for such inference.

However, small samples are tested whether this approach worked. The lack of time and computational power also impacts the data collection from Wikidata. The entity linking approaches are performed with the data from DBpedia. Future investigations are necessary to validate the kinds of conclusions that can be drawn from this study.

There is ample room for future research. We, therefore, propose future work to address as follows.

- Wider coverage of information: Other information provided by CH such as a person with significant control can provide us further insight and support broader analysis questions. In addition, other knowledge bases should be considered such as Crunchbase[50] and PermID[51].
- Further research could also be conducted with the financial filings and information in other countries to determine the effectiveness of the proposed approaches.
- A standardization of iXBRL/XBRL linked data vocabulary: Considerably more work will need to be done to determine the data model/ontology for both iXBRL or XBRL instances and all Linkbases of XBRL taxonomies. Having global standardization could be beneficial for the development of the financial ecosystem worldwide.
- Alternative entity linking approaches: Other information in the financial reports can be connected to knowledge bases such as accountants' names, addresses, and post codes. In addition, for a larger data set, alternative approaches for the linking may be necessary to achieve a high-quality ratio.
- Automatic transformation process: As the information from CH is updated on a frequent basis, a data pipeline (also known as an ETL pipeline) can be constructed to provide real-time information. In addition, a user interface or application may be developed to support the analysis and visualization tasks.

[50] https://www.crunchbase.com/.

[51] https://permid.org/.

References

Aidi, A., Mohd, H. & Mohd, N. (2019). Examining the Trend of the Research on eXtensible Business Reporting Language (XBRL): A Bibliometric Review. *International Journal of Innovation, Creativity and Change.* 5(2).

An Introduction to XBRL. (2021). XBRL. Retrieved April 20, 2021, from https://www.xbrl.org/the-standard/what/an-introduction-to-xbrl/.

Angelis, S., & Kotis, K. (2021). Generating and Exploiting Semantically Enriched, Integrated, Linked and Open Museum Data. In E. Garoufallou, M. A. Ovalle-Perandones and A. Vlachidis (Eds.), *Communications in computer and information science,* (pp. 367-379). Springer.

Ashraf, J., & Hussain, O. K. (2012). Integrating Financial Data Using Semantic Web for Improved Visibility. In *Proceedings of 8th International Conference on Semantics, Knowledge and Grids (SKG)* (pp. 265-268). Beijing, China: IEEE.

Asimadi, E., Reiff-Marganiec, S., Donnelly, B., Baker, J., and Fang, D. (2017). *Semantic approach to financial data integration for enabling new insights.*

Bao, J., Rong, G., Li, X., & Ding, L. (2010). Representing Financial Reports on the Semantic Web. In M. Dean, J. Hall, A. Rotolo, and S. Tabet (Eds.), *Lecture Notes in Computer Science, Semantic Web Rules* (pp. 144-152). Springer.

Berners-Lee, T. (2006). *Linked Data - Design Issues.* W3. https://www.w3.org/DesignIssues/LinkedData.html.

Bikakis, N., & Sellis, T. (2016). Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art. https://arxiv.org/pdf/1601.08059.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S. (2003). Adaptive name matching in information integration, *IEEE Intell. Syst.*, 18(5), 16–23. 10.1109/MIS.2003.1234765.

Build a Simple Map. (2021). Tableau. Retrieved June 25, 2021, from https://help.tableau.com/current/pro/desktop/en-us/maps_howto_simple.htm.

Build Maps in Tableau. (n.d). Tableau. Retrieved June 21, 2021, from https://help.tableau.com/current/pro/desktop/en-us/maps_build2.htm.

Burdick, D., Hernandez, M., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L., Stanoi, I. , Vaithyanathan, S. and Das, S. (2011). *Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study.*

Business population estimates for the UK and the regions 2020. (2020). Department for Business, Energy and Industrial Strategy. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/923565/2020_Business_Population_Estimates_for_the_UK_and_regions_Statistical_Release.pdf.

Carretié, H., Torvisco, B., & García, R. (2012). Using Semantic Web Technologies to Facilitate XBRL-based Financial Data Comparability.

Cohen, E. E. (2012). *Newcomer's Session*. XBRL. http://archive.xbrl.org/25th/sites/25thconference.xbrl.org/files/NewcomersSession_Yokohama2012_sm.pdf.

Companies register activities: 2019 to 2020. (2020). GOV.UK. https://www.gov.uk/government/statistics/companies-register-activities-statistical-release-2019-to-2020/companies-register-activities-2019-to-2020.

Companies House. (2020). GOV.UK. https://www.gov.uk/government/organisations/companies-house.

Company Reporting in the UK – an XBRL Success Story. (2015). XBRL UK. https://www.xbrl.org.uk/resources/whitepapers/UKcompanyReporting-XBRL-v1.pdf.

Dadzie, A. S., & Rowe, M. (2011). Approaches to visualising Linked Data: A survey, *SemanticWeb*, 2(2), 89–124. 10.3233/SW-2011-0037.

De Santo, A., and Holzer, A. (2020). Interacting with Linked Data: A Survey from the SIGCHI Perspective. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12).

Declerck, T. & Krieger, H. U. (2006). Translating XBRL Into Description Logic. An Approach Using Protege, Sesame and OWL.

Developer Guide - FRC Taxonomies. (2019, May 28). FRC. https://www.frc.org.uk/getattachment/917b5257-5279-4e65-8796-097bd2c1fba1/Final-Developer-Guide-FRC-Taxonomies-2019-08-13.pdf.

Dimou, A., Sande, M. V., Colpaert, P., Verborgh, R., Mannens, E. & Walle, R. (2014). *RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data.*

Dürst M., & Suignard M. (2005). *Internationalized Resource Identifiers (IRIs).* IETF. https://www.ietf.org/rfc/rfc3987.txt.

EHIEDU, & Chukwunweike, V. (2014). The impact of liquidity on profitability of some selected companies: The financial statement analysis (FSA) approach, *Research Journal of Finance and Accounting*, 5(5). https://core.ac.uk/download/pdf/234629826.pdf.

Elmagarmid, A. K., Ipeirotis, P. G & Verykios, V. S. (2007). Duplicate Record Detection: A Survey, *IEEE Trans. Knowl. Data Eng.*, 19(1), 1–16. 10.1109/TKDE.2007.250581.

Extensible Business Reporting Language (XBRL) 2.1. (2013). XBRL. http://www.xbrl.org/specification/XBRL-2.1/REC-2003-12-31/XBRL-2.1-REC-2003-12-31+corrected-errata-2013-02-20.html.

Facts About W3C. (2021). W3C. Retrieved May 3, 2021, from https://www.w3.org/Consortium/facts.

García, R., & Gil, R. (2010). Linking XBRL Financial Data. In Wood, D. (Ed.). *Linking Enterprise Data* (pp. 103-125). Springer US.

Gawriljuk, G., Harth, A., Knoblock, C.A. & Szekely, P. (2016). *A Scalable Approach to Incrementally Building Knowledge Graphs*.

Getting Started for Developers. (2021). XBRL. Retrieved May 26, 2021, from https://www.xbrl.org/the-standard/how/getting-started-for-developers/.

Hall, W., & O'Hara, K. (2009). Semantic Web. In Meyers R.A. (ed.), *Encyclopedia of Complexity and Systems* (pp. 8084-8104). Springer, New York. https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30440-3_478.

Hausenblas, M. (2012). *5-star Open Data.* 5-star Open Data. https://5stardata.info/en/.

Heyvaert, P., Chaves-Fraga, D., Priyatna, F., Corcho, O., Mannens, E., Verborgh, R., & Dimou, A. (2019). Conformance Test Cases for the RDF Mapping Language (RML). In B. Villazón-Terrazas and Y. Hidalgo-Delgado (Eds.), *Knowledge Graphs and Semantic Web. KGSWC 2019. Communications in Computer and Information Science* (pp. 162-173). Cham, Switzerland: Springer.

HM Revenue and Customs. (2021). GOV.UK. Retrieved May 25, 2021, from https://www.gov.uk/government/organisations/hm-revenue-customs.

Hoffman, C., & Rodríguez, M.M. (2013). Digitizing Financial Reports – Issues and Insights: A Viewpoint. *The International Journal of Digital Accounting Research*, 13, 73-98. 10.4192/1577-817/1577-8517-v13_3.

Hogan, A. (2014). *Linked Data and the Semantic Web Standards*. http://aidanhogan.com/docs/ldmgmt_semantic_web_linked_data.pdf.

Hogan, A. (2020). *The Web of Data*. SPRINGER NATURE.

Import map features from a file. (2021). Google. Retrieved June 21, 2021, from https://support.google.com/mymaps/answer/3024836?co=GENIE.Platform%3DDesktopandhl=en#zippy=%2Cstep-prepare-your-info.

Inline XBRL Part 0: Primer 1.1. (2015, December 09). XBRL. https://www.xbrl.org/WGN/inlineXBRL-part0/WGN-2015-12-09/inlineXBRL-part0-WGN-2015-12-09.html#sec-transformation.

Inline XBRL Part 1: Specification 1.1. (2013). XBRL. https://www.xbrl.org/specification/inlinexbrl-part1/rec-2013-11-18/inlinexbrl-part1-rec-2013-11-18.html.

iXBRL. (2021). XBRL. Retrieved May 21, 2021, from https://www.xbrl.org/the-standard/what/ixbrl/.

iXBRL Tagging Features. (2019, October 03). XBRL. https://www.xbrl.org/guidance/ixbrl-tagging-features/.

Janev, V., Graux, D., Jabeen, H., & Sallinger, E. (2020). *Knowledge Graphs and Big Data Processing.* Springer. https://link.springer.com/book/10.1007%2F978-3-030-53199-7.

Kämpgen, B., Weller, T., O'Riain, S., Weber, C. & Harth, A. (2014). Accepting the XBRL Challenge with Linked Data for Financial Data Integration. In V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, A. Tordai (Eds.), *The Semantic Web: Trends and Challenges* (pp. 595-610). Cham, Switzerland: Springer.

Lee, V., Goto, M., Hu, B., Naseer, A., Vandenbussche, P. Y., Shakair, G., & Rodrigues, E. (2014). Exploiting Linked Data in Financial Engineering. In K. Liu, S. R. Gulliver, W. Li, C. Yu (Eds.), *IFIP Advances in Information and Communication Technology* (pp. 116-125). Heidelberg: Springer.

Lee, V., Goto, M., & Izu, T. (2016). Identity Mapping Solution for Open Data Federation, *Fujitsu Scientific and Technical Journal*, 52(1), 61–72.

Li, H., & Zhai, J. (2016). Constructing Investment Open Data of Chinese Listed Companies Based on Linked Data. *Association for Computing Machinery*. 475-480. https://doi.org/10.1145/2912160.2912206.

Livieri, B., Zappatore, M., & Bochicchio, M. (2014). Towards an XBRL Ontology Extension for Management Accounting. In E. Yu, G. Dobbie, M. Jarke, S. Purao (Eds.), *Conceptual Modeling* (1st ed., pp. 289-296). Cham, Switzerland: Springer*Linked Data Cookbook.* (2014). W3C. Retrieved May 10, 2021, from https://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook.

Location Data that Tableau Supports for Building Map Views. (2021). Tableau. Retrieved June 25, 2021, from https://help.tableau.com/current/pro/desktop/en-us/maps_data.htm.

Mora-Rodriguez, M., Atemezing, G. A., & C. Preist. (2017). Adopting Semantic Technologies for Effective Corporate Transparency. In E. Blomqvist, D. Maynard, A. Gangemi, and A. Hoekstra (Eds.), The Semantic Web. (pp. 655-670). Portorož, Slovenia, Springer International Publishing AG.

Noy, N. & McGuinness, D. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. *Knowledge Systems Laboratory*. 32.

Ontologies. (2021). W3C. Retrieved May 3, 2021, from https://www.w3.org/standards/semanticweb/ontology.

O'Riain, S., Curry, E., & Harth, A. (2012). XBRL and open data for global financial ecosystems: A linked data approach. *International Journal of Accounting Information Systems*, 13 (2), 141–162. 10.1016/j.accinf.2012.02.002.

OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition). (2012). W3C. https://www.w3.org/TR/owl2-syntax/#Data_Properties.

OWL Web Ontology Language Guide. (2004). W3C. https://www.w3.org/TR/2004/REC-owl-guide-20040210/#owl_sameAs.

OWL Web Ontology Language Overview. (2019). W3C. Retrieved June 18, 2021, from https://www.w3.org/TR/owl-features/.

Pan, D., & Zhang, D. (2016). Research on XBRL Domain Ontology Construction. *Technology and Investment*, 07(01), 8-13, 10.4236/ti.2016.71002.

Piotroski, J. D. (2000). Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers, *Journal of Accounting Research*, 38, 1. 10.2307/2672906.

Prepare annual accounts for a private limited company. (2015). GOV.UK. https://www.gov.uk/annual-accounts/microentities-small-and-dormant-companies.

Radzimski, M., Sanchez-Cervantes, J.L., Garcia-Crespo, A., & Temiño-Aguirre, I. (2014). Intelligent Architecture for Comparative Analysis of Public Companies Using Semantics and XBRL Data. *International Journal of Software Engineering and Knowledge Engineering*, 24 (5), 801–823. https://doi.org/10.1142/S0218194014500314.

RDF. (2014, February 25). RDF Working Group. https://www.w3.org/RDF/.

RDF Schema 1.1. (2014). W3C. https://www.w3.org/TR/rdf-schema/.

RDF 1.1 Concepts and Abstract Syntax. (2014, February 25). W3C. https://www.w3.org/TR/rdf11-concepts/.

Reconciling. (2021). OpenRefine. Retrieved June 26, 2021, from https://docs.openrefine.org/manual/reconciling.

Resource Description Framework (RDF): Concepts and Abstract Syntax. (2004, February 10). W3C. https://www.w3.org/TR/rdf-concepts/.

Roohani, S. (2008). *What is the History of XBRL?*. XBRL Education. http://www.xbrleducation.com/edu/history.htm.

Roman, D., Alexiev, V., Paniagua, J., Elvesætera, B., von Zernichowa, B.M., Soylu, A., Simeonoy, B., & Taggar, C. The euBusinessGraph ontology: A light-weight ontology for harmonizing basic company information. (2021). SemanticWeb, 13(3), 1-28. 10.3233/SW-210424.

R2RML: RDB to RDF Mapping Language. (2012, September 27). https://www.w3.org/TR/r2rml/.

Sánchez-Cervantes, J. L., Alor-Hernández, G., Salas-Zárate, M. d. P., García-Alcaraz, J. L., & Rodríguez-Mazahua, L. (2018). FINALGRANT: A Financial Linked Data Graph Analysis and Recommendation Tool. In R. Valencia-García, M. A. Paredes-Valverde, M. d. P. Salas-Zárate, and G. Alor-Hernández (Eds.), *Studies in Computational Intelligence, Exploring Intelligent Decision Support Systems* (pp. 3-26). Springer.

Semantic Web. (2019). W3C. Retrieved May 3, 2021, from https://www.w3.org/standards/semanticweb/.

Shen, W., Wang, J., & Han J. (2015). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2), 443–460. 10.1109/TKDE.2014.2327028.

Skousen, C. J., Smith, K. R., & Wright, C. J. (2009). Detecting and predicting financial statement fraud: The effectiveness of the fraud triangle and SAS No. 99. In M. Hirschey, K. John, and A. K. Makhija (Eds.), *Advances in Financial Economics* (1st ed., pp. 53-81). Bingley, UK: Emerald JAI.

SPARQL Query Language for RDF. (2008). W3C. Retrieved June 9, 2021, from https://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.

SPARQL 1.1 Overview. (2013). W3C. Retrieved June 9, 2021, from https://www.w3.org/TR/sparql11-overview/.

Standard industrial classification of economic activities (SIC). (2008). GOV.UK. https://www.gov.uk/government/publications/standard-industrial-classification-of-economic-activities-sic.

The Standard for Reporting. (2018). XBRL. Retrieved April 20, 2021, from https://www.xbrl.org/the-standard/what/the-standard-for-reporting/.

The XBRL Standard. (2021). XBRL. Retrieved June 7, 2021, from https://specifications.xbrl.org/.

Tips and tricks for Google Maps on your computer. (2021). Google. Retrieved June 21, 2021, from https://support.google.com/maps/answer/6029919?hl=enand ref_topic=7001233.

Townley, G. (2018, August 21). GOV.UK. https://companieshouse.blog.gov.uk/2018/08/21/acronyms-and-initialisms-sic-codes/.

UK SIC 2007. (2022). *Office for National Statistics.* https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007.

Viola, F., Roffia, L., Antoniazzi, F., D'Elia, A., Aguzzi, C., & Salmon Cinotti, T. (2018). Interactive 3D Exploration of RDF Graphs through Semantic Planes, *Future Internet*, 10(8), 81. 10.3390/fi10080081.

Wang, W. L., Huang, M., & Wang, Y. (2014). Construction of XBRL Semantic Metamodel and Knowledge Base Based on Ontology, *Applied Machines and Materials*, 571-572, 1119–1128. 10.4028/www.scientific.net/AMM.571-572.1119.

Wenger, M. R., Thomas, M. A. & Babb, J. S. (2013). Financial reporting comparability: toward an XBRL ontology of the FASB/IFRS conceptual framework, *IJEF*, 7(1), 15. 10.1504/IJEF.2013.051754.

Wikidata: Tools/OpenRefine. (2021). OpenRefine. Retrieved June 26, 2021, from https://www.wikidata.org/wiki/Wikidata:Tools/OpenRefine?mscl-kid=8203e121bbee11eca381dd7c5cc8d8c2.

Wong, J. (2020*). String Matching With FuzzyWuzzy*. Towards Data Science. https://towardsdatascience.com/string-matching-with-fuzzywuzzy-e982c61f8a84.

XBRL guide for businesses. (2020, April 08). GOV.UK. https://www.gov.uk/government/publications/xbrl-guide-for-uk-businesses/xbrl-guide-for-uk-businesses.

XBLR in the UK. (2021). XBRL UK. Retrieved May 21, 2021, from https://www.xbrl.org.uk/projects/.

XBRL Project Directory. (2021). XBRL. Retrieved April 20, 2021, from https://www.xbrl.org/the-standard/why/xbrl-project-directory/.

XBRL Tagging Guide - FRC Taxonomies. (2019, August 13). FRC. https://www.frc.org.uk/getattachment/95337a3d-6ace-449b-a1f8-09906a5ffdf6/XBRL-Tagging-Guide-FRC-Taxonomies-2019-08-13.pdf.

XBRL Taxonomy Development Handbook. (2020). XBRL US. https://xbr-lus.github.io/docs/tdh.html.

Listing 1: Example of RDF transformation of iXBRL file

Panel A: iXBRL file – HTML format

```
<ns0:html…>…

<ix:nonNumeric contextRef="C_CA_CB"
name="http://xbrl.frc.org.uk/cd/2019-01-
01/business#NameEntityOfficer">Mr J D Millett</ix:nonNumeric>
…
<xbrli:context id="C_CA_CB">
   <xbrli:segment>
      <xbrldi:explicitMember
dimension="http://xbrl.frc.org.uk/cd/2019-01-
01/business#EntityOfficersDimension">http://xbrl.frc.org.uk/cd/2019-
01-01/business#Director1
      </xbrldi:explicitMember>
   </xbrli:segment>
     <xbrli:period>
         <xbrli:startDate>2019-01-01</xbrli:startDate>
         <xbrli:endDate>2019-12-31</xbrli:endDate>
     </xbrli:period>
</xbrli:context>

…</ns0:html>
```

Panel B: Expected RDF result

```
 …
[ rdf:type iXBRLInstanceOnto:NonNumeric, xbrll:Fact;
   xbrll:concept <http://xbrl.frc.org.uk/cd/2019-01-
01/business#NameEntityOfficer>;
   iXBRLInstanceOnto:hasContext <http://paul.ti.rw.
fau.de/~un62efef/thesis/fs_rdf/iXBRLFile_00634835#C_CA_CB> ;
xbrll:hasDimension [ xbrll:axis <http://xbrl.frc.org.uk/cd/2019-
01-01/business#EntityOfficersDimension>;
     xbrll:value <http://xbrl.frc.org.uk/cd/2019-01-
01/business#Director1>];
    xbrll:period [xbrll:endDate "2019-01-01"^^xsd:date ;
    xbrll:startDate "2019-12-31"^^xsd:date] ;
    xbrll:value "Mr J D Millett"],
  …
```

Panel C: RML transformation rule

```
<#factNonNumMapping>
 a rr:TriplesMap ;
 rml:logicalSource [
 rml:source "Prod223_2742_00677272_20191231.html";
 rml:iterator "//*[local-name()='nonNumeric']";
 rml:referenceFormulation ql:XPath ];

  rr:subjectMap [rr:termType rr:BlankNode ;
                 rr:class xbrll:Fact, iXBRLInstanceOnto:nonNumeric];
 rr:predicateObjectMap [rr:predicate xbrll:concept ;
                        rr:objectMap [ rml:reference "@name";
                                       rr:termType rr:IRI ] ];
 rr:predicateObjectMap [rr:predicate xbrll:period ;
 rr:objectMap [rr:parentTriplesMap  <#contextPeriodMapping>;
               rr:joinCondition [ rr:child "@contextRef" ;
                                  rr:parent "../@id" ] ] ];
…
<#contextPeriodMapping>
 a rr:TriplesMap ;
 rml:logicalSource [
 rml:source "Prod223_2742_00677272_20191231.html";
   rml:iterator "//*[local-name()='period']";
   rml:referenceFormulation ql:XPath ];

 rr:subjectMap [rr:termType rr:BlankNode ;
                rr:class xbrll:Period ];
 rr:predicateObjectMap [rr:predicate xbrll:startDate ;
                        rr:objectMap [ rml:reference "startDate";
                                       rr:datatype xsd:date ] ];
 rr:predicateObjectMap [rr:predicate xbrll:endDate ;
                        rr:objectMap [ rml:reference "endDate";
                                       rr:datatype xsd:date ] ].
…
```

Listing 2: Example of a fact after applying the transformation steps

Panel A: Fact of an officer with the link

```
  …
 [ rdf:type iXBRLInstanceOnto:NonNumeric, xbrll:Fact;
   xbrll:concept <http://xbrl.frc.org.uk/cd/2019-01-01/busi-
ness#NameEntityOfficer>;
   iXBRLInstanceOnto:hasDimensionAxis <http://xbrl.
frc.org.uk/cd/2019-01-01/business#EntityOfficers Dimension>;
    iXBRLInstanceOnto:hasDimensionValue
<http://xbrl.frc.org.uk/cd/2019-01-01/business# Director1>;
    schema:startDate "2019-01-01"^^xsd:date ;
    schema:endDate "2019-12-31"^^xsd:date ;
    xbrll:value "Mr J D Millett";
    rdfs:seeAlso <http://paul.ti.rw.fau.de/~un62efef/thesis/of-
ficer_profile.ttl#00634835_officer_MrJDMillett> ],
  …
```

Panel B: Numerical fact with the converted value

```
…
[ rdf:type iXBRLInstanceOnto:NonFraction,
  xbrll:Fact;
  xbrll:concept  <http://xbrl.frc.org.uk/fr/2019-01-
01/core#PropertyPlantEquipmentGrossCost> ;
  iXBRLInstanceOnto:hasFormat <http://www.xbrl.org/in-
lineXBRL/transformation/2010-04-20#numcommadot> ;
  iXBRLInstanceOnto:hasScale 0 ;
  xbrll:decimals 0 ;
  iXBRLInstanceOnto:hasUnitMeasure <http://www.
xbrl.org/2003/iso4217#GBP> ;
   iXBRLInstanceOnto:hasInstant "2018-12-31"^^xsd:date ;
   xbrll:value 1452392 ;
 iXBRLInstanceOnto:convertedValue 1452392.0 ],
…
```

Listing 3: Example of RDF transformation of XBRL taxonomy

Panel A: Expected RDF result

```
<http://xbrl.frc.org.uk/fr/2019-01-01/core#PropertyPlantEquipmentGrossCost>  rdf:type skos:Concept;
           rdfs:subClassOf xbrli:monetaryItemType, xbrli:debit, xbrli:instant;
           rdfs:label "Property, plant and equipment, gross / at cost"@en;
           dcterms:references  <core_PropertyPlantEquipmentGrossCost_refhttp://xbrl.frc.org.uk/general/ref/roles/FRS102>.

<core_PropertyPlantEquipmentGrossCost_refhttp://xbrl.frc.org.uk/general/ref/roles/FRS102>
           dcterms:references <http://xbrl.frc.org.uk/general/ref/roles/FRS102> ;
           dcterms:title "FRS";
           TaxonomyOnto:toNumber "102" .
           TaxonomyOnto:toParagraph "17.31.d" .
```

Panel B: RML transformation

```
<#elementMapping>
 a rr:TriplesMap;
  rml:logicalSource [
  rml:source "core-2019-01-01NS.xsd";
  rml:iterator "//*[local-name()='element']";
  rml:referenceFormulation ql:XPath  ];

 rr:subjectMap [rr:template "{//*[local-name()='schema']/@targetNamespace}#{@name}" ;
                rr:class skos:Concept ];

 rr:predicateObjectMap [rr:predicate TaxonomyOnto:hasType ;
                        rr:objectMap [ rml:reference "@type";
                                       rr:termType rr:IRI ] ];

 rr:predicateObjectMap [rr:predicate TaxonomyOnto:hasPeriod ;
                        rr:objectMap [ rml:reference "@periodType";
                                       rr:termType rr:IRI ] ];

 rr:predicateObjectMap [rr:predicate TaxonomyOnto:hasBalance ;
                        rr:objectMap [ rml:reference "@balance";
                                       rr:termType rr:IRI ] ].

<#referenceRoleRefMapping>
 a rr:TriplesMap;
   rml:logicalSource [
   rml:source "core-2019-01-01-reference.xml";
   rml:iterator "//*[local-name()='roleRef']";
   rml:referenceFormulation ql:XPath  ];

 rr:subjectMap [rml:reference "@roleURI" ;
                rr:termType rr:IRI ];

 rr:predicateObjectMap [rr:predicate TaxonomyOnto:hasRoleReference_reference;
                        rr:objectMap [ rr:template "{@href}" ] ].

<#arcRoleReferenceMapping>
 a rr:TriplesMap;
   rml:logicalSource [
   rml:source "core-2019-01-01-reference.xml";
   rml:iterator "//*[local-name()='referenceArc']";
   rml:referenceFormulation ql:XPath  ];
```

Panel C: RML transformation (continued)

```
 42  rr:subjectMap [rr:template "{@from}" ];
 43
 44  rr:predicateObjectMap [rr:predicate TaxonomyOnto:referredTo ;
 45                         rr:objectMap [ rr:template "{@to}" ] ].
 46
 47 <#ReferenceRoleMapping>
 48  a rr:TriplesMap;
 49   rml:logicalSource [
 50   rml:source "core-2019-01-01-reference.xml";
 51   rml:iterator "//*[local-name()='reference']";
 52   rml:referenceFormulation ql:XPath  ];
 53
 54  rr:subjectMap [rr:template "{@label}" ];
 55
 56  rr:predicateObjectMap [rr:predicate dcterms:references ;
 57                        rr:objectMap [ rr:template "{@label}{@role}" ] ].
 58
 59 <#ReferenceDetailMapping>
 60  a rr:TriplesMap;
 61   rml:logicalSource [
 62   rml:source "core-2019-01-01-reference.xml";
 63   rml:iterator "//*[local-name()='reference']";
 64   rml:referenceFormulation ql:XPath  ];
 65
 66  rr:subjectMap [rr:template  "{@label}{@role}" ];
 67
 68  rr:predicateObjectMap [rr:predicate dcterms:references ;
 69                         rr:objectMap [ rml:reference "@role" ;
 70                                        rr:termType rr:IRI  ] ];
 71
 72  rr:predicateObjectMap [rr:predicate <http://purl.org/dc/terms/title>;
 73                         rr:objectMap [ rml:reference "Name/." ] ];
 74
 75  rr:predicateObjectMap [rr:predicate TaxonomyOnto:toParagraph ;
 76                         rr:objectMap [ rml:reference "Paragraph/." ] ];
 77
 78  rr:predicateObjectMap [rr:predicate TaxonomyOnto:toNumber ;
 79                         rr:objectMap [ rml:reference "Number/." ] ].
```

Listing 4: Example of a concept in 2014 and 2019 FRC taxonomies

Panel A: Concept in 2014 FRC taxonomy

```
<http://xbrl.frc.org.uk/fr/2014-09-01/core#TaxTaxCreditOnProfitOr-
LossOnOrdinaryActivities> a skos:Concept;

   TaxonomyOnto:hasPeriod <http://paul.ti.rw.fau.de/~un62efef/the-
sis/ontology/general.ttl#duration>;

   TaxonomyOnto:hasType <<http://www.xbrl.org/dtr/type/non-nu-
meric#domainItemType>;

   rdfs:label "Held-to-maturity financial assets"@en.
```

Panel B: Concept in 2019 FRC taxonomy

```
<http://xbrl.frc.org.uk/fr/2019-01-01/core#TaxTaxCreditOnProfitOr-
LossOnOrdinaryActivities> a skos:Concept;

  TaxonomyOnto:hasPeriod <http://paul.ti.rw.fau.de/~un62efef/the-
sis/ontology/general.ttl#duration>;

  TaxonomyOnto:hasType <<http://www.xbrl.org/dtr/type/non-nu-
meric#domainItemType>.

  rdfs:label "Held-to-maturity investments, deferred tax (Depre-
cated 2019-01-01)"@en;

  rdfs:comment "This element has been deprecated but can still be
used by all FRS 102 and insurance filers"@en;

  dcterms:modified "2019-01-01"^^xsd:date.
```

Listing 5: Example of the proposed solution for FRC taxonomies in different years

Panel A: Concept in 2014 with unique elements

```
<http://xbrl.frc.org.uk/fr/2014-09-01/core#TaxTaxCreditOnProfitOr-
LossOnOrdinaryActivities> rdfs:label "Held-to-maturity financial as-
sets"@en;

  rdfs:seeAlso <http://example.org/core#TaxTaxCreditOnProfitOrLos-
sOnOrdinaryActivities>.
```

Panel B: Concept in 2019 with unique elements

```
<http://xbrl.frc.org.uk/fr/2019-01-01/core#TaxTaxCreditOnProfitOr-
LossOnOrdinaryActivities> rdfs:label "Held-to-maturity investments,
deferred tax (Deprecated 2019-01-01)"@en;

  rdfs:comment "This element has been deprecated but can still be
used by all FRS 102 and insurance filers"@en;

  dcterms:modified "2019-01-01"^^xsd:date;

  rdfs:seeAlso <http://example.org/core#TaxTaxCreditOnProfitOrLos-
sOnOrdinaryActivities>.
```

Panel C: Concept in both years with shared elements

```
<http://example.org/core#TaxTaxCreditOnProfitOrLossOnOrdinaryActivi-
ties> a skos:Concept;

   TaxonomyOnto:hasPeriod <http://paul.ti.rw.fau.de/~un62efef/the-
sis/ontology/general.ttl#duration>;

   TaxonomyOnto:hasType <<http://www.xbrl.org/dtr/type/non-nu-
meric#domainItemType>.
```

Listing 6: Example of officer name mapping by prefix similarity

Panel A: Officers in company 05499089

```
"05499089": [
   {"officer_name": "le bon yasmin",
    "split_name": ["le", "bon", "yasmin"]},

   {"officer_name": "lee tannaz",
    "split_name": ["lee", "tannaz"]},

   {"officer_name": "cook emma",
    "split_name": ["cook","emma"]},
  …]
```

Panel B: Mapping ratio by officer

Name in iXBRL Report	Compared Name in Companies House	Score
y	le	0.0
	bon	0.0
	yasmin	0.166
le	le	1.0
	bon	0.0
	yasmin	0.0
bon	le	0.0
	bon	1.0
	yasmin	0.0
["y", "le", "bon"]	**["le", "bon", "yasmin"]**	**2.166**
y	lee	0.0
	tannaz	0.0
le	lee	0.667
	tannaz	0.0
bon	lee	0.0
	tannaz	0.0
["y", "le", "bon"]	**["lee", "tannaz"]**	**0.667**
y	cook	0.0
	emma	0.0
le	cook	0.0
	emma	0.0
bon	cook	0.0
	emma	0.0
["y", "le", "bon"]	**["cook","emma"]**	**0.0**

Listing 7: Query result for financial analysis questions

```
PREFIX ex: <http://example.org#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>


<http://paul.ti.rw.fau.de/~un62efef/thesis/fs_rdf/iXBRL-
File_09785763.ttl>
        ex:hasCurrency                  "Pound sterling" ;
        ex:hasTurnoverY18               8970536.0 ;
        ex:hasTurnoverY19               8686221.0 ;
        ex:hasAverageTurnover19_18      8828378.5 ;
        ex:hasRevenueGrowthrateY19_18   -3.16943157019825794100 ;
        ex:hasGrossProfitY18            3003859.0 ;
        ex:hasGrossProfitY19            3062444.0 ;
        ex:hasGrossMarginY18            33.486 ;
        ex:hasGrossMarginY19            35.256 ;
        ex:hasCurrentAssetY18           2749344.0 ;
        ex:hasCurrentAssetY19           2495327.0 ;
        ex:hasCurrentLiabilityY18       2272750.0 ;
        ex:hasCurrentLiabilityY19       2609138.0 ;
        ex:hasCurrentRatioY18           1.21 ;
        ex:hasCurrentRatioY19           0.956.


<http://paul.ti.rw.fau.de/~un62efef/thesis/fs_rdf/iXBRL-
File_03831189.ttl>
        ex:hasCurrency                  "Pound sterling" ;
        ex:hasTurnoverY18               14329764.0 ;
        ex:hasTurnoverY19               26621856.0 ;
        ex:hasAverageTurnover19_18      20475810.0 ;
        ex:hasRevenueGrowthrateY19_18   85.78014264575466839500;
        ex:hasGrossProfitY18            6392557.0 ;
        ex:hasGrossProfitY19            4468573.0 ;
        ex:hasGrossMarginY18            44.61;
        ex:hasGrossMarginY19            16.785 ;
        ex:hasCurrentAssetY18           6319760.0 ;
        ex:hasCurrentAssetY19           11388124.0 ;
        ex:hasCurrentLiabilityY18       3244028.0 ;
        ex:hasCurrentLiabilityY19       7016844.0 ;
        ex:hasCurrentRatioY18           1.948 ;
        ex:hasCurrentRatioY19           1.623.
```

Appendix A – FRC taxonomy file structure (Developer Guide – FRC Taxonomies, 2019, pp. 31-33)

Location	Prefix	Comments and namespaces
xbrl.frc.org.uk/		
cd/		Folder containing Common Data Taxonomy
yyyy-mm-dd/		
business/		
bus-yyyy-mm-dd.xsd	bus	Basic schema for Business Taxonomy http://xbrl.frc.org.uk/cd/yyyy-mm-dd/business
bus-yyyy-mm-dd-definition.xml bus-yyyy-mm-dd-presentation.xml bus-yyyy-mm-dd-label.xml bus-yyyy-mm-dd-reference.xml		
bus-full-yyyy-mm-dd.xsd	bus-full	Entry point for Full Business Taxonomy http://xbrl.frc.org.uk/cd/yyyy-mm-dd/business-full
countries-regions/		
countries-yyyy-mm-dd.xsd	countries	Basic schema for Countries and Regions Taxonomy http://xbrl.frc.org.uk/cd/yyyy-mm-dd/countries
countries-yyyy-mm-dd-definition.xml countries-yyyy-mm-dd-presentation.xml countries-yyyy-mm-dd-label.xml countries-yyyy-mm-dd-reference.xml		
countries-full-yyyy-mm-dd.xsd	countries-full	Entry point for Full Countries and Regions Taxonomy http://xbrl.frc.org.uk/cd/yyyy-mm-dd/countries-full
currencies/		
currencies-yyyy-mm-dd.xsd	curr	Basic schema for Currencies taxonomy http://xbrl.frc.org.uk/cd/yyyy-mm-dd/currencies
currencies-yyyy-mm-dd-definition.xml currencies-yyyy-mm-dd-presentation.xml currencies-yyyy-mm-dd-label.xml currencies-yyyy-mm-dd-reference.xml		
currencies-full-yyyy-mm-dd.xsd	currencies-full	Entry point for Full Currencies Taxonomy http://xbrl.frc.org.uk/cd/yyyy-mm-dd/currencies-full
languages/		
languages-yyyy-mm-dd.xsd	lang	Basic schema for Languages Taxonomy http://xbrl.frc.org.uk/cd/yyyy-mm-dd/languages
languages-yyyy-mm-dd-definition.xml languages-yyyy-mm-dd-presentation.xml languages-yyyy-mm-dd-label.xml languages-yyyy-mm-dd-reference.xml		
languages-full-yyyy-mm-dd.xsd	languages-full	Entry point for Full Languages Taxonomy http://xbrl.frc.org.uk/cd/yyyy-mm-dd/languages-full
fr/		Folder containing Core Financial Reporting Taxonomy
yyyy-mm-dd/		
core/		
frc-core-yyyy-mm-dd.xsd	core	Basic schema for Core Financial Reporting Taxonomy http://xbrl.frc.org.uk/fr/yyyy-mm-dd/core
frc-core-yyyy-mm-dd-definition.xml frc-core-yyyy-mm-dd-presentation.xml frc-core-yyyy-mm-dd-label.xml frc-core-yyyy-mm-dd-reference.xml		
frc-core-full-yyyy-mm-dd.xsd	core-full	Entry point for Full Core Financial Reporting Taxonomy http://xbrl.frc.org.uk/fr/yyyy-mm-dd/core-full
FRS-101		Folder containing FRS 101 Taxonomy
yyyy-mm-dd/		
FRS-101-yyyy-mm-dd.xsd	FRS-101	Schema for FRS 101 Taxonomy http://xbrl.frc.org.uk/FRS-101/yyyy-mm-dd
FRS-101-yyyy-mm-dd-presentation.xml		
FRS-102		Folder containing FRS 102 Taxonomy
yyyy-mm-dd/		
FRS-102-yyyy-mm-dd.xsd	FRS-102	Schema for FRS 102 Taxonomy http://xbrl.frc.org.uk/FRS-102/yyyy-mm-dd
FRS-102-yyyy-mm-dd-definition.xml FRS-102-yyyy-mm-dd-presentation.xml		
IFRS		Folder containing UK IFRS Full Taxonomy
yyyy-mm-dd/		
IFRS-yyyy-mm-dd.xsd	IFRS	Schema for UK IFRS Taxonomy http://xbrl.frc.org.uk/FRS/yyyy-mm-dd
IFRS-yyyy-mm-dd-presentation.xml		
general/		Folder containing taxonomies with basic definitions
yyyy-mm-dd/		
common/		
common-yyyy-mm-dd.xsd	common	Schema defining common tags using in all taxonomies http://xbrl.frc.org.uk/general/yyyy-mm-dd/common
common-yyyy-mm-dd-label.xml		
ref/		
ref-yyyy-mm-dd.xsd	uk-ref	Schema defining additional reference parts used in UK http://xbrl.frc.org.uk/general/yyyy-mm-dd/ref
types/		
types-yyyy-mm-dd.xsd	types	Schema defining additional data types used in UK http://xbrl.frc.org.uk/general/yyyy-mm-dd/types

Appendix A – FRC taxonomy file structure (Developer Guide – FRC Taxonomies, 2019, pp. 31-33) (continued)

reports/		Folder containing reports taxonomies
yyyy-mm-dd/		
accrep/		
accrep-yyyy-mm-dd.xsd	accrep	Basic schema for Accountant's Report Taxonomy http://xbrl.frc.org.uk/reports/yyyy-mm-dd/accrep
accrep-yyyy-mm-dd-definition.xml accrep-yyyy-mm-dd-presentation.xml accrep-yyyy-mm-dd-label.xml accrep-yyyy-mm-dd-reference.xml		
accrep-full-yyyy-mm-dd.xsd		Entry point for Accountant's Report Taxonomy http://xbrl.frc.org.uk/fr/yyyy-mm-dd/core-full
aurep/		
aurep-yyyy-mm-dd.xsd	aurep	Basic schema for Audit Report Taxonomy http://xbrl.frc.org.uk/reports/yyyy-mm-dd/aurep
aurep-yyyy-mm-dd-definition.xml aurep-yyyy-mm-dd-presentation.xml aurep-yyyy-mm-dd-label.xml aurep-yyyy-mm-dd-reference.xml		
aurep-fees-2014-01-01.xsd	aurep-fees	Entry point for fees section of Audit Report Taxonomy http://xbrl.frc.org.uk/reports/yyyy-mm-dd/aurep-fees
aurep-fees-2014-01-01-presentation.xml		
aurep-full-yyyy-mm-dd.xsd	aurep-full	Entry point for Full Audit Report Taxonomy http://xbrl.frc.org.uk/reports/yyyy-mm-dd/aurep-full
direp/		
direp-yyyy-mm-dd.xsd	direp	Basic schema for Directors' Report Taxonomy http://xbrl.frc.org.uk/reports/yyyy-mm-dd/direp
direp-yyyy-mm-dd-definition.xml direp-yyyy-mm-dd-presentation.xml direp-yyyy-mm-dd-label.xml direp-yyyy-mm-dd-reference.xml		
direp-full-yyyy-mm-dd.xsd	direp-full	Entry point for full Directors' Report Taxonomy http://xbrl.frc.org.uk/reports/yyyy-mm-dd/direp-full
direp-income-2014-01-01.xsd	direp-income	Entry point for income section of Directors' Report Taxonomy http://xbrl.frc.org.uk/reports/yyyy-mm-dd/direp-income
direp-income-2014-01-01-presentation.xml		

Appendix B – SPARQL query for financial analysis

```
PREFIX iXBRLInstanceOnto: <http://paul.ti.rw.fau.de/~un62efef/thesis/ontology/iXBRLInstanceOnto.ttl#>
PREFIX xbrll: <https://w3id.org/vocab/xbrll#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ex: <http://example.org/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <http://schema.org/>

CONSTRUCT { ?comp ex:hasTurnoverY19 ?val_R;
                  ex:hasTurnoverY18 ?valprev_R;
                  ex:hasAverageTurnover19_18 ?avg_R;
                  ex:hasRevenueGrowthrateY19_18 ?growth_R;
                  ex:hasCurrency ?unit_label;
                  ex:hasGrossProfitY19 ?val_G; ex:hasGrossProfitY18 ?valprev_G;
                  ex:hasGrossMarginY19 ?grossmargin; ex:hasGrossMarginY18 ?grossmargin_prev;
                  ex:hasCurrentAssetY19 ?val_CA; ex:hasCurrentAssetY18 ?valprev_CA;
                  ex:hasCurrentLiabilityY19 ?val_CL; ex:hasCurrentLiabilityY18 ?valprev_CL;
                  ex:hasCurrentRatioY19 ?currentRatio; ex:hasCurrentRatioY18 ?currentRatio_prev
}WHERE {
      ?comp  iXBRLInstanceOnto:hasFact  [xbrll:concept ?turnover;
                                         iXBRLInstanceOnto:hasUnitMeasure ?unit;
                                         schema:endDate ?endDate_prev;
                                         iXBRLInstanceOnto:convertedValue ?valprev_R].
      OPTIONAL{?unit rdfs:label ?unit_label.}
      ?comp  iXBRLInstanceOnto:hasFact  [xbrll:concept ?turnover;
schema:endDate ?endDate;
                                          iXBRLInstanceOnto:convertedValue ?val_R].

  FILTER ((?endDate_prev > "2018-06-30"^^xsd:date && ?endDate_prev <= "2019-06-30"^^xsd:date) &&
          (?endDate > "2019-06-30"^^xsd:date && ?endDate <= "2020-06-30"^^xsd:date))

      OPTIONAL { ?comp  iXBRLInstanceOnto:hasFact [xbrll:concept ?GP;
schema:endDate ?endDate_prev;
                                                   iXBRLInstanceOnto:convertedValue ?valprev_G].}
      OPTIONAL { ?comp  iXBRLInstanceOnto:hasFact [xbrll:concept ?GP;
schema:endDate  ?endDate;
                                                   iXBRLInstanceOnto:convertedValue ?val_G].}
      OPTIONAL { ?comp  iXBRLInstanceOnto:hasFact [xbrll:concept ?CA;
                                                   iXBRLInstanceOnto:has-
Instant ?endDate_prev;
                                                   iXBRLInstanceOnto:convertedValue ?valprev_CA].}
      OPTIONAL { ?comp  iXBRLInstanceOnto:hasFact [xbrll:concept ?CA;
                                                   iXBRLInstanceOnto:has-
Instant ?endDate;
                                                   iXBRLInstanceOnto:convertedValue ?val_CA].}
      OPTIONAL { ?comp  iXBRLInstanceOnto:hasFact [xbrll:concept ?creditor;
                                                   iXBRLInstanceOnto:hasDimensionAxis ?group;
                                                   iXBRLInstanceOnto:hasDimensionValue ?CF;
                                                   iXBRLInstanceOnto:has-
Instant ?endDate_prev;
                                                   iXBRLInstanceOnto:convertedValue ?valprev_CL].}
      OPTIONAL { ?comp  iXBRLInstanceOnto:hasFact [xbrll:concept ?creditor;
                                                   iXBRLInstanceOnto:hasDimensionAxis ?group;
                                                   iXBRLInstanceOnto:hasDimensionValue ?CF;
```

Appendix B – SPARQL query for financial analysis (continued)

```
VALUES ?turnover {<http://paul.ti.rw.fau.de/~un62efefhesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2014-09-01/core#TurnoverRevenue>

  <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2019-01-01/core#TurnoverRevenue>

  <core:TurnoverRevenue> }

 VALUES ?GP {<http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2014-09-01/core#GrossProfitLoss>

  <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2019-01-01/core#GrossProfitLoss>

  <core:GrossProfitLoss>}

 VALUES ?CA { <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2014-09-01/core#CurrentAssets>

  <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2019-01-01/core#CurrentAssets>

  <core:CurrentAssets>}

 VALUES ?creditor { <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2014-09-01/core#Creditors>

  <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2019-01-01/core#Creditors>

  <core:Creditors>}

 VALUES ?CF {<http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2014-09-01/core#CurrentFinancialInstruments>

  <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2019-01-01/core#CurrentFinancialInstruments>

  <core:CurrentFinancialInstruments>}


FILTER (?group != <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/cd/2014-09-01/business#GroupCompanyDataDimension> ||

       ?group != <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/cd/2019-01-01/business#GroupCompanyDataDimension> ||

      ?group !=  <business:GroupCompanyDataDimension>)


#Calculation  filter > 0  to solve divide by zero

  FILTER (?valprev_R > 0)

   BIND ((?val_R + ?valprev_R)/2 AS ?avg_R)

   BIND ((((?val_R - ?valprev_R)/?valprev_R)*100000)/1000 AS ?growth_R)

   BIND (ROUND(((?valprev_G)/?valprev_R)*100000)/1000 AS ?grossmargin_prev)

  FILTER (?val_R > 0)

   BIND (ROUND(((?val_G)/?val_R)*100000)/1000 AS ?grossmargin)

  FILTER (?val_CA > 0)

   BIND (ROUND(((?val_CA)/(?val_CL)) AS ?currentRatio)

  FILTER (?valprev_CA > 0)

   BIND (ROUND(((?valprev_CA)/(?valprev_CL)) AS ?currentRatio_prev)

}
```

Appendix C – SPARQL query for industry analysis

```
#1 How many companies have an asset less and more than GBP 10,000 in 2019?

PREFIX iXBRLInstanceOnto: <http://paul.ti.rw.fau.de/~un62efef/thesis/ontology/iXBRLInstanceOnto.ttl#>

PREFIX xbrll: <https://w3id.org/vocab/xbrll#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>


SELECT  ?value (COUNT(DISTINCT ?instance) AS ?Totalcompanies)

WHERE {

          ?instance iXBRLInstanceOnto:hasFact [xbrll:concept ?concept;

                                               iXBRLInstanceOnto:hasInstant ?date;

                                               iXBRLInstanceOnto:convertedValue ?amount]


VALUES ?concept{ <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2014-09-01/core#FixedAssets>

 <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2019-01-01/core#FixedAssets>

 <core:FixedAssets>

}


  FILTER(?date > "2018-12-31"^^xsd:date && ?date <= "2019-12-31"^^xsd:date).

  BIND(IF (?amount < 10000, 'assets < 10000',

       IF(?amount >= 10000, 'assets >= 10000','0')) AS ?value)

} GROUP BY ?value


#2 What is the profit of companies registered in SIC code 68209?

PREFIX iXBRLInstanceOnto: <http://paul.ti.rw.fau.de/~un62efef/thesis/ontology/iXBRLInstanceOnto.ttl#>

PREFIX xbrll: <https://w3id.org/vocab/xbrll#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

PREFIX rov: <http://www.w3.org/ns/regorg#>

PREFIX sic: <http://business.data.gov.uk/companies/def/sic-2007/>

PREFIX schema: <http://schema.org/>

PREFIX ns5: <https://spec.edmcouncil.org/fibo/ontology/FND/Arrangements/Reporting/>


  SELECT (AVG(?amount) AS ?avg)

WHERE {


          ?report ns5:isSubmittedBy ?comp.


          ?report iXBRLInstanceOnto:hasFact [xbrll:concept ?Profit;

                                             schema:endDate ?endDate;

                                             iXBRLInstanceOnto:convertedValue ?amount].

          ?comp_1 owl:sameAs ?comp;

                  rov:orgActivity sic:68209.


VALUES ?Profit { <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2014-09-01/core#ProfitLoss>
<http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2019-01-01/core#ProfitLoss>

 <core:ProfitLoss>}


} ORDER BY ?report
```

Appendix C – SPARQL query for industry analysis (continued)

```
#3 What are the top 5 companies which have the largest average number of employees by
industries (SIC code) in 2019?

PREFIX iXBRLInstanceOnto: <http://paul.ti.rw.fau.de/~un62efef/thesis/ontology/iXBRLIn-
stanceOnto.ttl#>

PREFIX xbrll: <https://w3id.org/vocab/xbrll#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

PREFIX rov: <http://www.w3.org/ns/regorg#>

PREFIX sic: <http://business.data.gov.uk/companies/def/sic-2007/>

PREFIX ns5: <https://spec.edmcouncil.org/fibo/ontology/FND/Arrangements/Reporting/>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

PREFIX schema: <http://schema.org/>


SELECT ?sic ?BusinessActv (AVG(?avgoutput) AS ?avg)

WHERE {

     ?report ns5:isSubmittedBy ?comp.

     ?report  iXBRLInstanceOnto:hasFact [xbrll:concept ?concept;

                                         schema:endDate ?endDate;

                                         xbrll:value ?AvgEmployeeNo].


      ?comp_1 owl:sameAs ?comp;

                  rov:orgActivity ?sic.


      ?sic skos:prefLabel ?BusinessActv;

            sic:condensedNotation ?SICCode.


 FILTER(?endDate > "2019-06-30"^^xsd:date && ?endDate <= "2020-06-30"^^xsd:date)

 BIND(IF(?AvgEmployeeNo = '-',"0"^^xsd:integer,

      IF(?AvgEmployeeNo = ' 0 ', "0"^^xsd:integer,

        IF(?AvgEmployeeNo = '0', "0"^^xsd:integer, xsd:integer(?AvgEmployeeNo) )))
AS ?avgoutput) .


VALUES ?concept{ <http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2014-09-01/core#AverageNumberEmployeesDuringPeriod>
<http://paul.ti.rw.fau.de/~un62efef/thesis/taxonomy/taxonomy_allcon-
cept.ttl#http://xbrl.frc.org.uk/fr/2019-01-01/core#AverageNumberEmployeesDuringPeriod>

<core:AverageNumberEmployeesDuringPeriod> }


} GROUP BY ?sic ?BusinessActv

ORDER BY DESC(?avg)

LIMIT 5
```

Appendix D – SPARQL query for company analysis

```
#1 Which companies have parent and/or subsidiary companies?
PREFIX CompanyProfileOnto: <http://paul.ti.rw.fau.de/~un62efef/thesis/ontology/CompanyProfileOnto.ttl#>
PREFIX schema: <http://schema.org/>

CONSTRUCT {
    ?company schema:parentOrganization ?parentlink
}
WHERE {
    ?company CompanyProfileOnto:hasCompanyNumber_ParentCompany ?parentnum;
             CompanyProfileOnto:hasLinks_ParentCompany ?parentlink;
             schema:parentOrganization ?parentname.
}

#2 What is the full name of Mr J D Millet? Is there any further information that can we know about this person e.g., birthday, nationality, occupation, and address?
PREFIX xbrll: <https://w3id.org/vocab/xbrll#>
PREFIX iXBRLInstanceOnto: <http://paul.ti.rw.fau.de/~un62efef/thesis/ontology/iXBRLInstanceOnto.ttl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

CONSTRUCT {
    ?officer ?p   ?o.
    ?o     ?p1    ?o1.
    ?o1    ?p2    ?o2.
}
WHERE {
          ?instance iXBRLInstanceOnto:hasFact [rdfs:seeAlso ?link;
                                               xbrll:value "Mr J D Millett"].

          ?officer owl:sameAs ?link;
                   ?p   ?o.
          OPTIONAL {?o     ?p1    ?o1.
                    ?o1    ?p2    ?o2.}
}
```

Section B.4

How to visualize relationships between supervisory board and management board members and auditors using Neo4j

(with. Julia Vetter)

Working Paper

Presented at:

Nodes Conference 2021, online

Accepted for presentation at:

39th Eurasia Business and Economics Society Conference 2022,

Rome, Italy

2nd International Conference on Accounting, Auditing and Finance

Hangzhou, China

Contents – Section B.4

How to visualize relationships between supervisory board and management board members and auditors using Neo4j

Abstract

This paper illustrates how relationships between supervisory board and management board members and auditors can be visualized using the graph database Neo4j. The sample includes DAX30 companies in the year 2019. The topic is of interest for two reasons. First, the special institutional setting follows a so called “principle of separation”. The supervision and the management of corporations should be clearly separated. In the past and the present, there are various examples of multiple mandates and personal ties between the members. Second, those relationships are not only difficult to identify, but also to analyze. Therefore, we have chosen a graph database (Neo4j) to detect, visualize and examine personal ties.

Keywords

Corporate Governance; Two-Tier System; Germany; Relationships; Neo4j

1 Introduction

This paper uses Neo4j to analyze relationships between supervisory and management board members of companies listed in the German stock index (DAX30).

The topic is of particular interest for various reasons. First, personal ties between supervisory and management board members and their auditors can be seen critical as they might create an elite network. Consequently, their objectivity and independence in performing tasks might be impaired. Hence, supervisory board members and auditors might not exercise their oversight tasks properly and be reluctant to raise critical issues.

Second, it is difficult to analyze and visualize the personal network of DAX30 supervisory and management board members and their auditors. It is extremely time-consuming to manually compare and transfer the CVs of 480 supervisory board members, 197 management board members and 56 auditors into a database. In addition, it is difficult to analyze the connections between the persons of interest. Established software such as MS Excel provide support for this, but do not have a suitable tool for analyzing the connections. In addition, connections cannot be visualized either. However, it is difficult for people without IT skills to find a suitable alternative.
Therefore, it is interesting to investigate how to use the graph database Neo4j for the analysis and visualization of the relationships between those board members as well as their auditors.

This paper is organized as follows. First, the problem of analyzing and visualizing huge amounts of data is explained in chapter two. Thereby, different types of databases and the graph database Neo4j are presented. The third chapter describes the unique institutional setting in Germany and the underlying data. The results for DAX30 companies, limited to the top management level and a certain backward time window are then presented in chapter four. Chapter five summarizes the main results and outlines future research avenues.

2 Basic knowledge

2.1 Use of data

Large amounts of data offer a variety of difficulties. Not only does the processing of the data become more difficult with increasing data size, but it also complicates the recognition of correlations. In the case of management boards and supervisory boards, these relationships exist between the individual members of the management and supervisory boards. For this paper, all the information found on members of the management board and supervisory board were gathered in a database.

A database system is required to store, process and retrieve a large amount of data. In concrete terms, a data management system, also known as a database management system (DBMS), is a software system that uses a standard method to store and organize data. In order to enter, retrieve, modify and generally work with the program, the user executes a specific query language, also called a manipulation language (Sumathi & Esakkirajan, 2007).

The DBMS determines the underlying database model depending on the form in which the data is loaded and how the relationships between them are mapped. There are several types of DBMS (Sumathi & Esakkirajan, 2007). There are different types of databases, such as network databases, relational databases and NoSQL databases (Sumathi & Esakkirajan, 2007). The latter ones can be divided into key value databases, column databases, document databases and graph databases (Abramova & Bernardino, 2013, Indrawan-Santiago, 2012). NoSQL databases are much faster at handling very large amounts of data than traditional relational databases. Unlike other databases, NoSQL databases do not use the standard tabular relationships that relational databases use. Instead, you can query and store data in NoSQL databases by a variety of other means, depending on the software you are using. Neo4j belongs to graph databases. Figure 1 shows several types of database management systems (Abramova & Bernardino, 2013; Indrawan-Santiago, 2012).

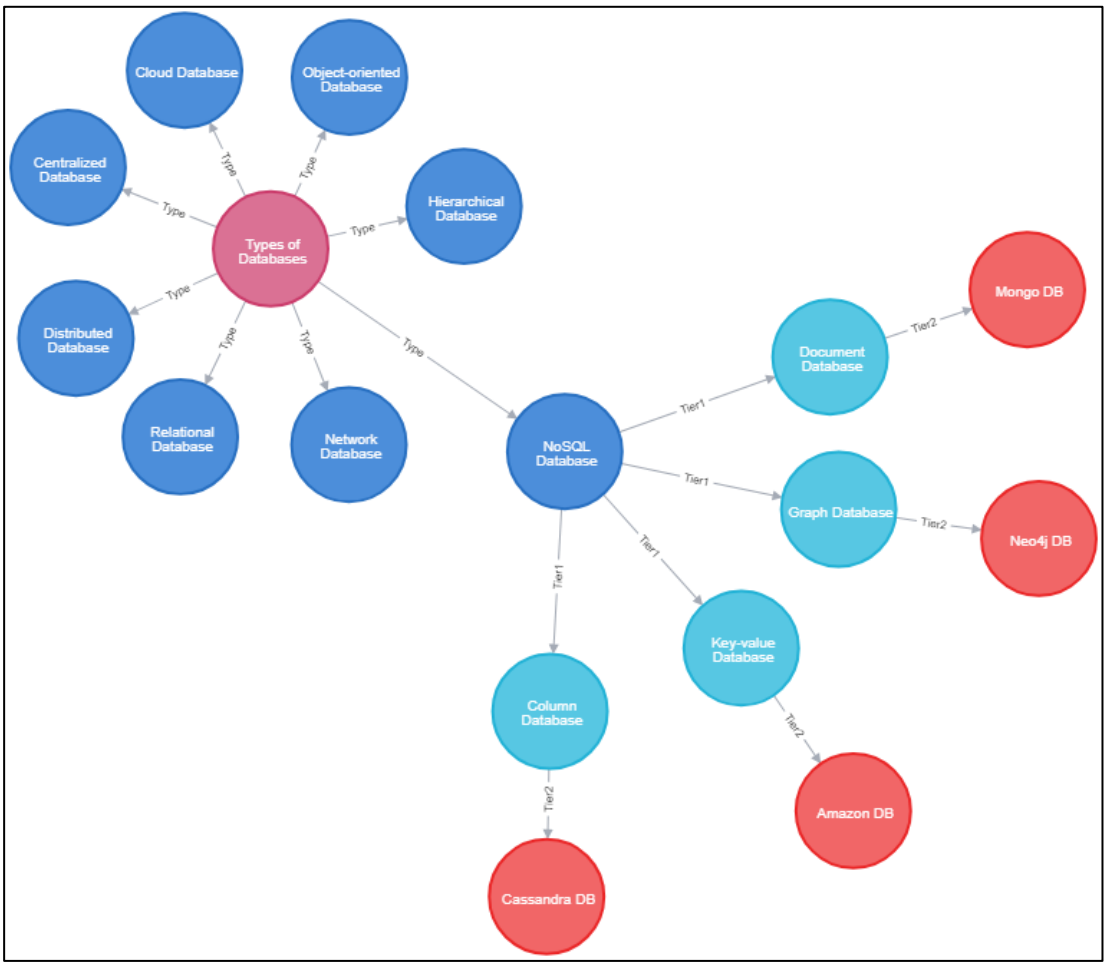


Figure 1: Types of database management systems

The figure also includes other types of databases, such as object-oriented databases, which are less common and represent a combination of data storage and object-oriented programming, and hierarchical databases, which are now considered obsolete due to their low flexibility as well as network databases. Ultimately, databases are to be differentiated in terms of the type of storage, not only in terms of their logical structures, in which cloud databases, centralized databases and distributed databases can be distinguished (Domdouzis, 2021).

2.2 Literature Review

Traditionally, relational databases have been the primary storage structure in most data management and data retrieval applications. Relational databases organize data in tables defined by sets of rows and columns. Each table represents an entity type while columns represent its attributes and rows can be considered instances of that type (Vicknair et al., 2010).

Additionally, there may be logical connections among different tables. The connections are defined by specifying the unique primary keys of the relevant tables. More specifically, the one-to-many relationship is realized by migrating one's own key to foreign tables, and the many-to-many relationships are demonstrated by creating additional key tables that contain the primary keys of the other two entities. Retrievals in relational databases are usually performed using Structural Query Language (SQL).

One of the limitations of relational models is that if the data contains a large number of relationships, huge joins of large tables are required.

These large data problems, such as social network modeling, bioinformatics computations, and building information modeling (BIM) data management are becoming more common in academia and industry today. Storing, retrieving and manipulating such complex data becomes very costly when using traditional relational database system approaches (Miller, 2013) because of the frequent misleading migration of keys and numerous key tables. Relational databases work with primary keys and foreign keys that relate the individual tables to each other. The larger the number of tables in the database, the slower the performance in relation to the response to queries (Hernández et al., 2016). With the increasing demand for storing large amounts of interconnected data, new storage alternatives to relational databases were developed. These new systems are categorized as NoSQL systems in which graphs are one of the most optimized databases for interconnected data (Batra & Tyagi 2012; Zhang, 2017).

2.1.2 Graph database

Graph databases belong to the category of NoSQL databases. A graph database is a database in which the data is displayed and stored using a graph. A graph consists of nodes and edges, which represent the relationships between the nodes. The nodes and edges can have additional attributes. In addition, each edge has a property that represents the relationship. The design of graph structure allows navigating and filtering through correlations and patterns. Moreover, the highly dynamic data model, in which all nodes are connected by relations, allows for fast traversals along the edges between vertices (Miller, 2013).

Graph databases specialize in storing networked information and traversing it efficiently. In this case, the so-called nodes and relationships are used to create the visualization.

This helps to uncover and visualize hidden relationships without additional effort. This becomes clear when observing the following simple example of a property graph:

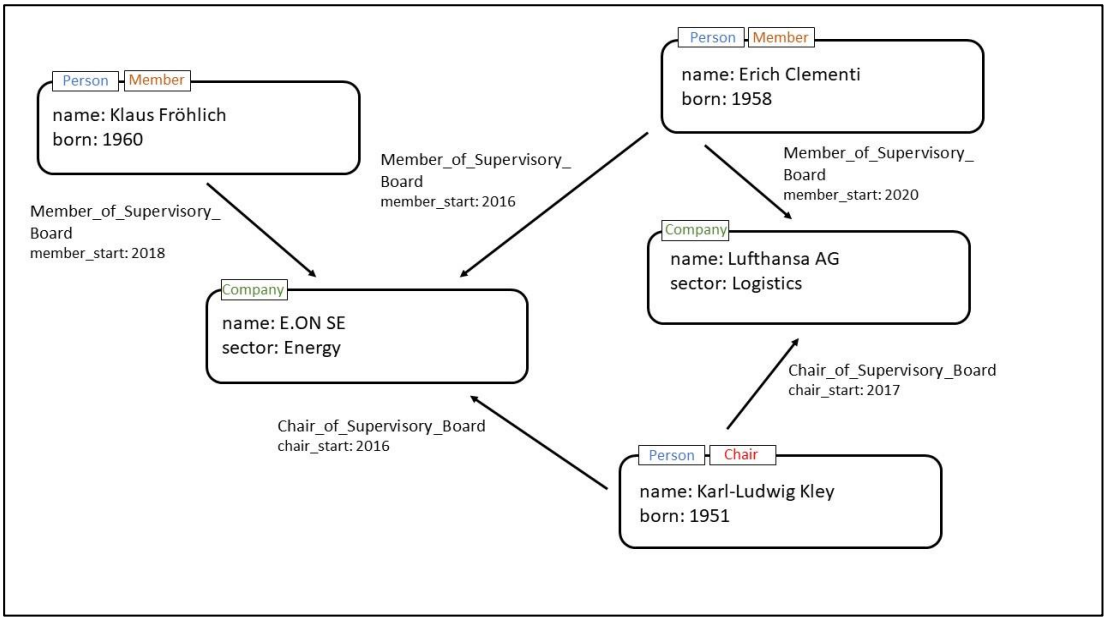


Figure 2: Property graph example

The nodes are visualized as circles connected by edges, which are the lines. Each node represents a data set. In the diagram, the left node contains two attributes or properties, name and age as well as the two labels ‘Person’ and ‘Member’, which are used to categorize the node and will play an important role in the query afterwards. The edges are the connecting link between the data objects. In most cases, the established relationships are directed from one object to the other.

In addition, the edges can be given a label and a property just like the nodes. The label is used to correctly assign the relationship, the property in turn equips the relationship with a characteristic (Paul et al., 2019; Robinson et al., 2013).

The example in Figure 2 illustrates the approach once again. The two nodes Karl-Ludwig Kley and Lufthansa AG are connected by an edge. It is labeled as “Chair_of_Supervisory_Board” and as an attribute it shows the “chair_start: 2017”, i.e. the year in which the person has been elected as the chair of the supervisory board.

In certain use cases and in the presence of strongly networked data, the graph databases offer various advantages. Firstly, the query speed is not dependent on the data volume.

Even with sophisticated command queries, the duration of the processing remains unchanged and effective execution is still possible. Secondly, this database model is characterized by its comfortable structure, which is not visible in relational databases due to the constant need to switch back and forth between tables.

Moreover, there are no rigid or inflexible table structures. The graphical representation allows a simple illustration of the relationships. Furthermore, the principle according to which this graph model works is more intuitive and easier to understand.

2.1.3 Graph database vs. relational database

The advantages of a graph database are obvious in most cases. The data is directly networked, stored and processed. This structure makes it possible to avoid complex queries such as recursively nested joins and this leads to an efficient traversal.

Traversing a graph means that the graph is traversed starting from a start node. Usually, a further node is searched for, such as the way from one city to another. Due to the efficient traversing, the performance is much higher than that of a relational database. This is especially true for systems that are specialized in searching for routes, since the structure of a graph database is designed exactly for such cases. A relational database would have to rely on complex queries as recursively nested joins, which are very performance-intensive (Hernández et al., 2016).

The scaling of a graph database, however, is not as easy as with other NoSQL databases. If the number of nodes and edges in the graph is too large for a single server, partitioning is used to split the graph so that it is distributed over several systems. Nonetheless, it is not always easy to find an appropriate place for partitioning in the graph. Even with an equally distributed relevance of all nodes in the graph, there is no mathematically exact method to minimize the number of intersected edges (Paul et al., 2019).

However, since a relational database also has difficulties with scaling, this is not a direct disadvantage compared to a relational database but compared to other NoSQL databases. A big disadvantage of a graph database is the query language. So far, there is no common query language due to the lack of standardization between graph databases. There are various query languages and graph models. (Vicknair et al. 2010; Cheng et al., 2019; Gong et al., 2018).

3 Data & institutional setting

In the following, the visualization of relationships between board members in Neo4j is presented by means of some example cases.

The following section is organized as follows. First, the data for our examples is presented considering the German two-tier system and the legal framework in Germany is illustrated.

The institutional setting of Germany is unique and it is important to understand the institutional background before considering the examples. The next subsection explains the "Deutschland AG". The last two chapters present examples of multiple mandates and personal ties which are illustrated in Neo4j.

3.1 Data

We base our examination of German companies listed in the DAX30 in the year 2019. Only those members of the supervisory board or management board who still held their office at the end of the 2019 financial year are considered. Changes during the year (e.g. board member retirements) are not taken into account.

Company	**Supervisory board**	**Management board**	**Auditors**
Adidas	16	6	2
Allianz	12	10	2
BASF	12	7	2
BMW	20	8	2
Bayer	20	7	2
Beiersdorf	12	8	2
Continental	20	8	2
Covestro	12	4	2
Daimler	20	8	2
Deutsche Bank	19	7	2
Deutsche Börse	16	6	2
Deutsche Post	20	8	2
Telekom	20	9	2
EON	20	5	2
Fresenius Medical Care	6	7	2

Fresenius	15	7	2
HeidelbergCement	12	7	2
Henkel	16	5	2
Infineon	16	4	2
Linde[375]	11	8	-
Lufthansa	21	6	2
Merck	16	5	2
MTU	12	4	2
Munich Re	20	9	2
RWE	20	2	2
SAP	18	8	2
Siemens	20	8	2
VW	20	8	2
Vonovia	12	4	2
Wirecard[376]	6	4	-
Sum	**480**	**197**	**56**

Table 1: Summary statistics

For each person, a variety of data was collected, such as year and place of birth, current position, secondary employment as well as the last three professional positions. The information is hand-collected from a variety of sources.

The amount of information about the people of interests varies. Some companies offer a detailed CV in PDF format. Other companies, however, provide only a text as a brief description. In principle, we assume that the information on the companies' websites is correct. Additionally, further information on the persons of interest derives from the internet. Suitable sources include social networks such as LinkedIn, but also news articles. Since the information on LinkedIn usually comes from the people themselves, it can be assumed that the information is also correct.

[375] It is not possible to find detailed information of the executives and auditors of Linde plc.
[376] We do not have any information about the auditors of Wirecard because the annual report for the year 2019 was not published.

3.2 German two-tier system

Since the structure of boards of directors in Germany, and thus their supervision, is very different from other countries, Germany offers a unique environment for investigating multiple board mandates. It provides a two-tier structure in which the management and supervision of the company is not to be carried out by the same body, but by two separate administrative bodies, the supervisory board and the management board.

The management board is responsible for conducting the firm's business according to Section 76 (1) German Stock Corporation Act (GSCA). Its members are appointed by the supervisory board (Section 84 (1) GSCA) which is responsible for supervising and advising the management board (Section 111 (1) GSCA). The supervisory board is elected by the shareholders at the shareholders' meeting (Section 101 (1) GSCA).

In accordance with Section 105 GSCA, members of the supervisory board may not at the same time be members of the management board.

Depending on the number of employees, the supervisory board of many listed companies must further include employee representatives (§ 4 One Third Participation Act, § 7 Co-determination Act). The supervisory board thus represents the interests of the owners and the employees.

3.3 Legal framework

In Germany, it is possible to be a member of several supervisory boards at the same time. The number of mandates is regulated by the German Stock Corporation Act (GSCA), which limits the number of mandates to ten. Within this framework, chairmanships are counted as two mandates. In order to comply with the principle of separation, members of the management board cannot be members of the supervisory board at the same time (Section 105 (1) GSCA).

The German Corporate Governance Code (GCGC), constituting soft law, is supplementary to the legal requirements of the GSCA. Section C.4 of the GCGC suggests a maximum of five supervisory board mandates for "normal" members and a maximum of two supervisory board mandates for members of the management board. [377]

[377] Note that chair positions count twice in this context.

In addition, there should be no more than two former management board members in the supervisory board. In addition, Section C.7 of the GCGC recommends that more than half of the shareholder representatives should be independent.[378]

3.4 The "Deutschland AG"

The so-called "Deutschland AG"[379] describes a personnel network of supervisory and management board members of German DAX30 companies (Reckendrees, 2013).

The network has dissolved continuously over the years due to stricter regulations. However, the historical connections between supervisory and management board members are still of interest as personal ties might impair the required independence of the management and supervisory board members (Gröls, 2011; Oehmichen, 2011).

4 Results

As already mentioned, a large amount of information was collected on the members of the supervisory board and the board of management. Our analysis focuses on a variety of examples. Through these, the advantages and exemplary use of Neo4j are shown. The exemplary examples can also be applied to a different or larger database.

In our examples, the nodes have different colors and present, among others, companies, supervisory and management boards as well as their members.

[378] Section C.7 of the GCGC includes among others the following criteria for the assessment of a member's independence: no personal or business relationships with the firm or its management board that may provoke conflicts of interest, no membership in the supervisory board of the same firm in the two years prior to the appointment to the supervisory board, no family relationship with a member of the management board.

[379] Like the *old boy network* in English.

For all examples shown below, the respective colors have the following meaning:

Color	Meaning
Orange	Firm
Green	Management Board
Light Blue	Member of the management board
Dark Blue	Chair of the management board
Red	Supervisory board
Light Pink	Member of the supervisory board
Dark Pink	Chair of the supervisory board
Beige	Field of studies
Light Yellow	University
Dark Green	PhD
Yellow	Auditor (person)
Purple	Auditor (occupation)

Table 2: Different colors and their definition

4.1 Multiple mandates

In the following, examples of multiple mandates of German DAX supervisory board members are illustrated with the help of Neo4j. In principle, it is possible to be a member of more than one supervisory board in Germany. For this reason, it is interesting to see who is on how many supervisory boards and on which ones.

Figure 3 presents the supervisory boards of the five German DAX companies Deutsche Post AG, Munich Re, Lufthansa AG, E.ON SE and BMW AG as well as their members. In addition to that, the graph also shows the management board of Lufthansa AG.

Figure 3: Multiple mandates

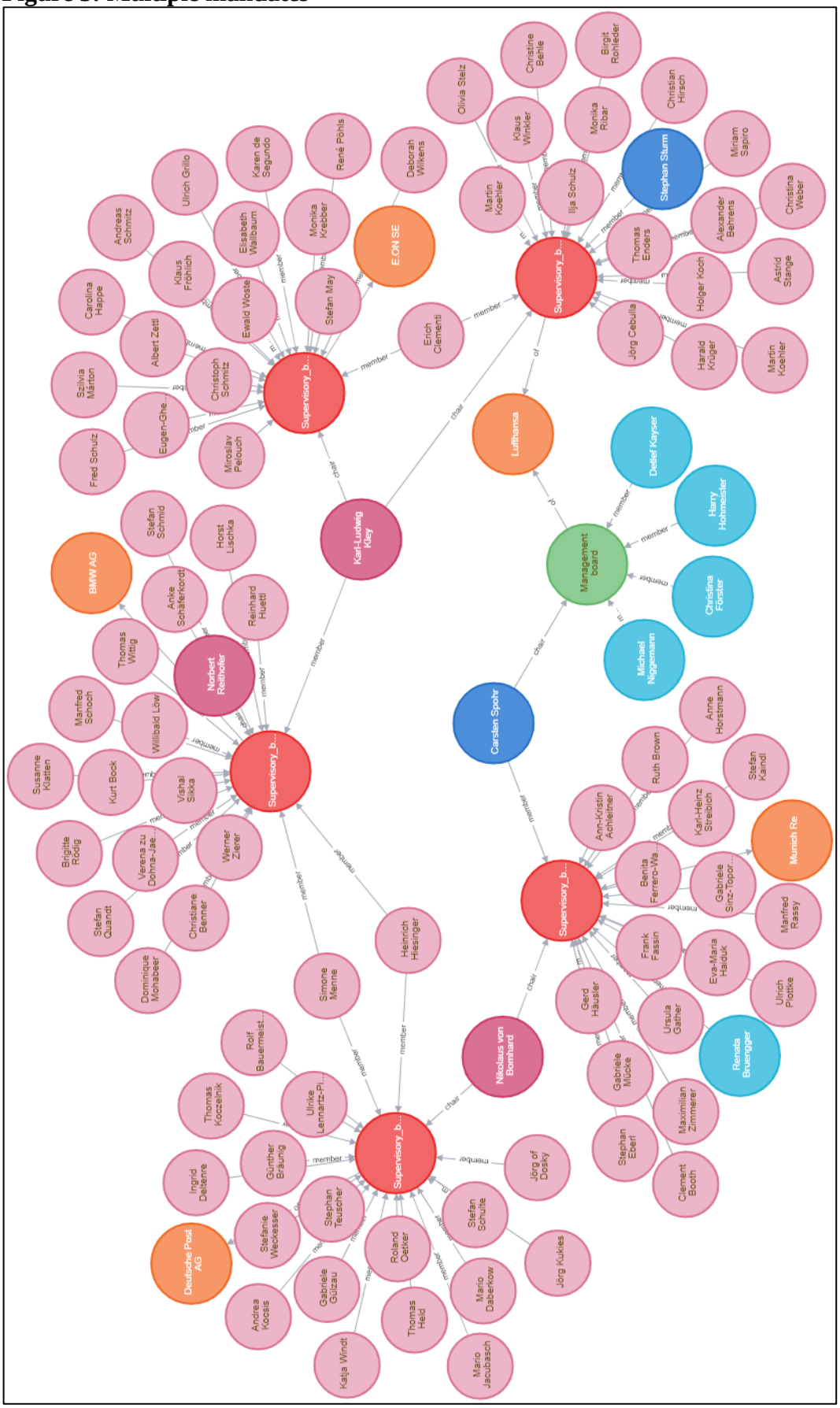


Figure 3 reveals that Karl-Ludwig Kley was the chair of the supervisory board of E.ON SE and Lufthansa AG and a member of the supervisory board of BMW AG in 2019.

In addition, Heinrich Hiesinger and Simone Menne were members of the supervisory boards of BMW AG and Deutsche Post AG. The supervisory board chair of Deutsche Post AG, Nikolaus von Bromhard, was also the chair of the supervisory board of Munich Re. In addition, Carsten Spohr, the supervisory board chair of Munich Re was simultaneously chair of the management board at Lufthansa.

The example illustrates that there are several personal connections between the supervisory and management boards of these five DAX companies.

Neo4j helps to visualize the relationships of multiple mandates. With the help of the Neo4j, one can easily detect if a person has a seat on multiple supervisory boards at the same time.

4.2 Personal ties

Multiple mandates are just one possibility to detect if there are relationships between the supervisory and management boards of German DAX companies. The personal ties between management and supervisory board members constitute another interesting aspect. In order to detect personal ties, a variety of information such as university, highest degree and former positions was hand-collected. A personal tie between two persons consists if their paths have crossed in the past (e.g., same university, same firm) at the same time.[380]

Figure 4 depicts personal ties between two members of the management boards of Daimler and Fresenius: Martin Daum (Daimler) and Stephan Sturm (Fresenius). In addition to his work at the management board of Daimler, Martin Daum is also a member of the supervisory board of Lufthansa. Both members studied at the University of Mannheim at the same time. Hence, Martin Daum and Stephan Sturm may know each other from their studies.

[380] Personal ties are a first hint that the two persons might actually know each other. However, we cannot rule out the possibility that the persons do not know each other.

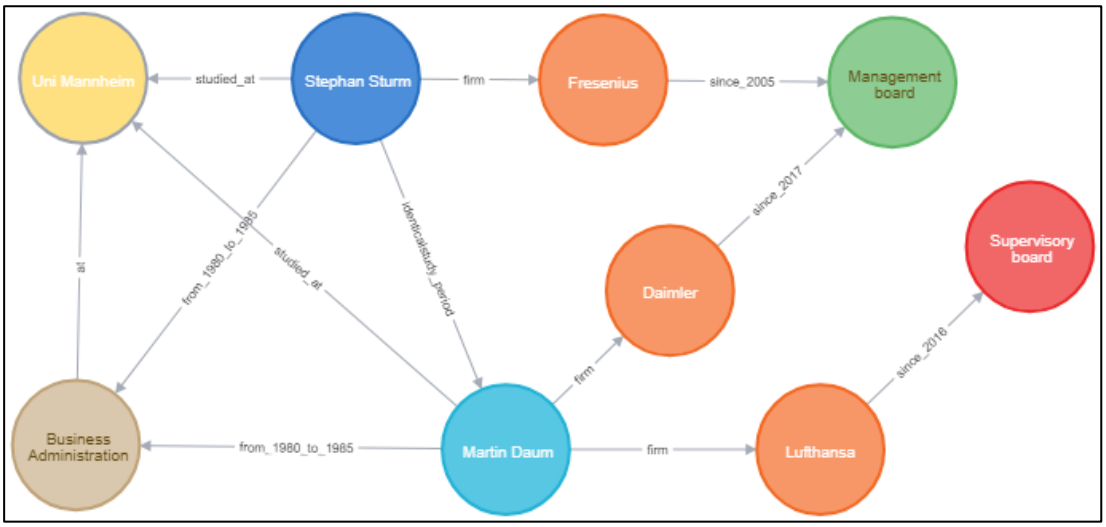


Figure 4 Personal ties

Figure 5 depicts another example for personal ties. Two members of the management board of Munich Re were at the same university at the same time: Ludger Arnoldussen and Timotheus Höttges. Both studied business administration at the University of Cologne from 1983 until 1988. Timotheus Höttges is also a member of the supervisory board of Daimler and Henkel.

Interestingly, Munich Re is audited by KPMG. The responsible auditor is Frank Ellenbürger who also studied business administration at the University of Cologne from 1982 until 1987. In addition, both Ludger Arnoldussen and Frank Ellenbürger hold a doctoral degree from the University of Cologne.

Hence, two members of the management board of Munich Re, Ludger Arnoldussen and Timotheus Höttges might know the auditor of Munich Re, Frank Ellenbürger, since both were at the University of Cologne at the same time. Both did not only study at the same time but also stayed at university afterwards to obtain a PhD. The example shows that the "Deutschland AG" has not completely dissolved yet. In addition, personal ties might impair the independence of supervisory and management board members.

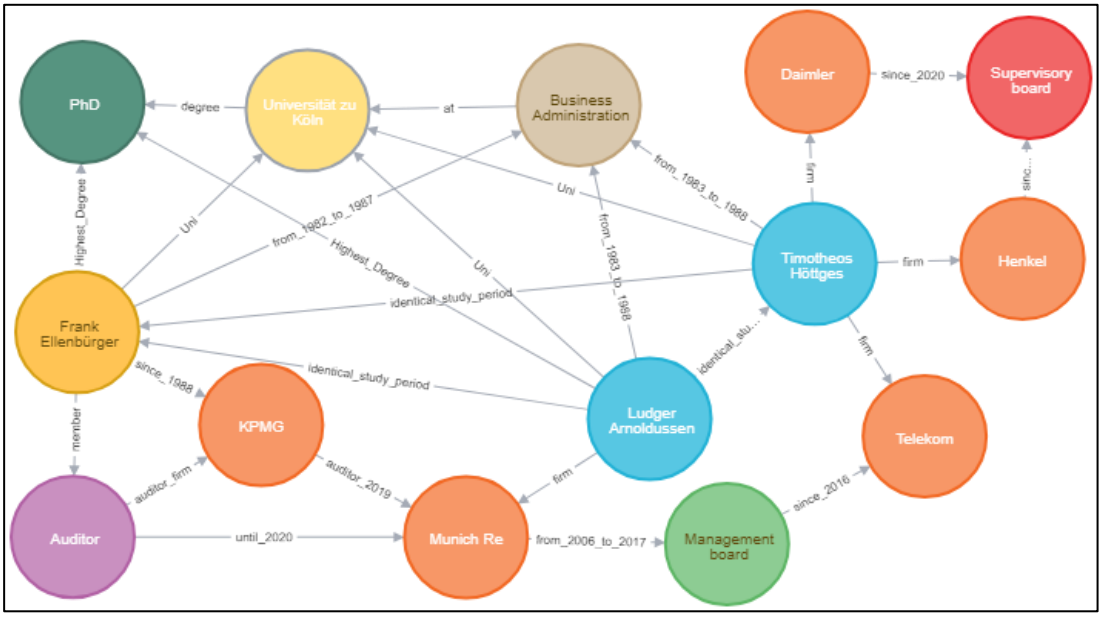


Figure 5: Personal ties with auditors

5 Conclusion

This paper shows exemplarily how Neo4j can be used to visualize personal networks in the German two-tier system. The results reveal that the members of the management and supervisory boards have various personal networks. This might cause problems because the objectivity and independence of supervisory board members and auditors could be impaired, which in turn affects their oversight ability in a negative way.

The analysis reveals that various persons of interest are members in several supervisory boards at the same time. In addition, it can be assumed that members of different management boards know each other because they attended the same university at the same time. Moreover, in one case, a member of the management board is personally connected with the auditor of the firm as both studied at the same university simultaneously.

With this paper, it has been shown that the analysis with Neo4j is significantly simplified. In the future, Neo4j could be used to analyze the members of the supervisory board and the management board of the entire DAX30. Furthermore, the same methodology could be applied in another institutional setting, e.g. the one-tier system in the U.S. or the UK.

Future research could also focus on creating a knowledge graph using RDF data. This is technically more demanding and much more difficult to implement for non-IT-savvy people. Nevertheless, interesting evaluations could be carried out independently of Neo4j.

References

Abramova, V., & Bernardino, J. (2013). NoSQL databases: MongoDB vs Cassandra. Proceedings of the international Conference on computer science and software engineering.

Batra, S., & Tyagi, C. (2012). Comparative analysis of relational and graph databases. *International Journal of Soft Computing and Engineering* 2(2). 509-512.

Cheng, Y., Ding, P., Wang, T., Lu, W., & Du, X. (2019). Which Category Is Better: Benchmarking Relational and Graph Database Management Systems. *Data Science and Engineering*, 4(4), 309-322.

Domdouzis, K. (2021). *Concise Guide to Databases: A Practical Introduction*. Springer International Publishing.

Gong, F., Ma, Y., Gong, W., Li, X., Li, C., & Yuan, X. (2018). Neo4j graph database realizes efficient storage performance of oilfield ontology. *PLOS ONE*. 13(11).

Gröls, M. (2011). Die letzten Herren der "Deutschland AG"? *Der Aufsichtsrat*: (07-08): 106-107.

Hernández, D., Hogan, A., Riveros, C., Rojas, C., & Zerega, E. (2016). Querying wikidata: Comparing sparql, relational and graph databases. *International Semantic Web Conference*, 88-103.

Indrawan-Santiago, M. (2012). Database Research: Are We at a Crossroad? Reflection on NoSQL. 15th International Conference on Network-Based Information Systems. IEEE, 45-51.

Miller, J. J. (2013). Graph database applications and concepts with Neo4j. Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA, 2324(36).

Oehmichen, J. (2011). Mehrfachmandate von Aufsichtsratsmitgliedern: Eine Panel-Analyse ihrer Wirkung in deutschen Unternehmen. In Lindstädt (Ed.), *Schriften zu Management, Organisation und Information*, Augsburg, Germany: Rainer Hampp Verlag.

Paul, S., Mitra, A., & Koner, C. (2019). A Review on Graph Database and its representation. *International Conference on Recent Advances in Energy-efficient Computing and Communication*, 1-5.

Reckendrees, A. (2013), Historische Wurzeln der Deutschland AG. In A. Ahrens, B. Gehlen & A. Reckendrees (Eds.) *Die „Deutschland AG“* (pp. 57-84). Klartext Verlag.

Robinson, I., Webber, J., & Eifrem, E. (2013). *Graph databases*. O'Reilly Media, Inc.

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., & Wilkins, D. (2010). A comparison of a graph database and a relational database: a data provenance perspective. *Proceedings of the 48th annual southeast regional conference*. 1-6.

Sumathi, S., & Esakkirajan, S. (2007). Fundamentals of relational database management systems, Springer, 47.

Zhang, Z. J. (2017). Graph databases for knowledge management. *IT Professional*, 19(6), 26-32.

Section B.5

What can we learn from Knowledge Graphs?

A Wirecard perspective

Working Paper

Presented at:

Knowledge Graph Conference 2021,

New York, USA

Current Issues in Business and Economic Studies Conference 2022

London, United Kingdom

Accepted for presentation at:

Accounting Education SIG Annual Conference 2022,

Glasgow, Scotland

Contents – Section B.5

What can we learn from Knowledge Graphs?

A Wirecard perspective

Abstract

The bankruptcy of German financial services company Wirecard AG, resulting from accounting irregularities and lack of oversight, amplified calls for more extensive financial supervision. German laws require members of management boards, supervisory boards and auditors to be ‘independent’ from each other. However, this only focuses on the current state of affairs and does not include long-term relationships. In order to capture these aspects, we created a knowledge graph containing details of each person of interest, such as previous roles and education. In order to increase the scope of this information, the database was enriched with data from external open graph databases. The goal of this paper is to analyze interrelations between persons of interest and explore knowledge graphs as a tool in the domain of accounting and auditing.

Keywords

Wirecard; Knowledge Graph; Linked Data; Auditing; Corporate Governance

1 Introduction

Considering the accounting irregularities and lack of oversight of the German financial services company Wirecard AG, which ultimately ended in insolvency and arrests of CEO Markus Braun, the calls for more extensive financial supervision amplified (Krahnen & Langenbucher, 2020; Véron, 2020). Journalists, as well as academics were dumbfounded how an organization listed among Germany's 30 largest publicly traded company in its blue-chip stock market index DAX (Deutscher Aktienindex) could deceive shareholders, financial regulators and the public eye for such a long time (Alderman & Schuetze, 2020; Traufetter et al., 2020).

The downfall of what was considered to be one of the most promising tech companies was the leading incentivewith the motivation to explore ways of improving accountability and transparency in the regulatory domain. We identified that the most essential aspect of German laws, which requires members of the management board, members of the supervisory board and auditors to be 'independent' from each other, focuses mainly on current affairs. However, long-term relationships among top executives could date all the way back to shared university studies, working together in the same company or being member of the same organizations.

One way of structuring this information is a so-called knowledge graph, a term coined by Google in 2012 (Singhal, 2012). While common relational databases store information in tables (relations) with columns (attributes) and rows (records), knowledge graphs use a graph structure. Nodes represent real world entities and resources, while edges describe their interrelation (Ehrlinger & Wöß, 2016). The strength of a knowledge graph lies in its capability to enrich the limited information in the database through links to external open graph databases available on the internet. As knowledge graphs emphasize relationships between entities, they allow for traversal and visualizations of relationships.

The aim of this paper is to use knowledge graphs to investigate the people behind the Wirecard scandal and ultimately understand whether there are differences in the composition and members of the supervisory board and management board that provides reason for further conclusions. The resulting findings will also be compared with the members of the supervisory boards and management boards of the other DAX30 companies.

The second chapter gives an overview of the theoretical background. First, the institutional setting is described with the german two-tier-system, the legal framework and the story behind the Wirecard scandal.

The third chapter focuses on the data and the research methodology. The data part gives an overview of the data origin, data scope and data linking. The research methodology focuses on the creation of the knowledge graph, querying the knowledge graph and the types of visualization.

The fourth chapter contains the analyses. Here, the focus is on presenting the situation of the management board and supervisory board at Wirecard. It then shows how knowledge graphs can be used to analyze the links between the members of the management board and supervisory board. In particular, industry connections and interboard connections are explored.

Finally, the fifth chapter concludes the analysis.

2 Institutional setting & Wirecard story

This chapter provides an overview of the theoretical foundations utilized in this paper. First, a short description of the principles and legal basis in German public corporations is outlined. The second chapter provides an overview of Wirecard and how it came to collapse. The role of KPMG is also discussed. Afterwards, the basics of linked data and knowledge graphs are explained.

2.1 Institutional setting

The institutional setting of Germany is unique and it is thus important to understand the institutional background before considering the examples.. The last two chapters then present examples of multiple mandates and personal ties which are visualized. It is important to note that this overview does not claim to be comprehensive, but rather is intended to provide a basic understanding necessary for further description.

German two-tier system

Germany provides a unique setting to investigate multiple board mandates as the board structure differs a lot from the board structure in the U.S. and other countries. In Germany, corporate boards have a so-called two-tier board structure.

The two-tier system follows a principle of separation, which implies that two separate administrative bodies are charged with the management and supervision of the firm: the management board and the supervisory board. The management board is responsible for conducting the firm's business (Section 76 (1) GSCA). Its members are appointed by the supervisory board which is responsible for supervising and advising the management board (Section 84 (1) GSCA). The supervisory board is elected by the shareholders at the shareholders' meeting (Section 101 (1) GSCA).

Depending on the number of employees, the supervisory board of listed companies must further include employee representatives (§ 4 One Third Participation Act, § 7 Co-determination Act). The supervisory board therefore represents the owners' and employees' interests. The principle of separation is reflected by the requirement that members of the supervisory board are not allowed to be part of the management board at the same time (Section 105 (1) GSCA).

Legal framework

It is possible to be a member of various supervisory boards at the same time. However, the German Stock Corporation Act (GSCA) limits the maximum number to ten mandates, whereas chair mandates count as two mandates. Members of the management board are not allowed to be members of the supervisory board as this violates the principle of separation (Section 111 (1) GSCA).

The German Corporate Governance Code (GCGC), which constitutes soft law, complements the legal requirements of the GSCA. Section C.4 of the GCGC recommends a maximum of five supervisory mandates for "normal" members[1] and a maximum of two supervisory board members for management board members. In addition, there should be no more than two former management board members in the supervisory board. In addition, Section C.7 of the GCGC recommends that more than half of the shareholder representatives should be independent.[2]

[1] Note that chair positions count twice in this context.

[2] Section C.7 of the GCGC includes among others the following criteria for the assessment of a me ber's independence: no personal or business relationships with the firm or its management board that may provoke conflicts of interest, no membership in the supervisory board of the same firm in the two years prior to the appointment to the supervisory board, no family relationship with a member of the management board.

2.2 Wirecard story

Business model and history of Wirecard

Wirecard AG was a payment service provider based in Aschheim near Munich. The business model was the technical processing of electronic payments, with a percentage retained as a fee for each payment made. Payment processing was also mostly extended to include risk management and fraud prevention (Peemöller et al., 2020). Transaction volume in 2018 was around 125 billion euros, according to Wirecard (Lenz, 2020). Wirecard was reported to have 313.000 customers (Bartz et al., 2020).

The stock corporation was founded in 1999 as Wire Card and got into trouble when the dotcom bubble burst (Rasch, 2020). After important data was lost in a burglary and Wire Card lost its technological edge as a result, insolvency had to be filed for the first time (Velte & Graewe, 2021). Laptops of Markus Braun and Jan Marsalek were lost in the burglary. Wirecard's founder suspected that an insider had been responsible, with the aim of bringing about the insolvency (Bartz et al., 2020).

In January 2002, shortly after filing for bankruptcy, the company was acquired by EBS Holding and merged with the latter under the name Wire Card. Markus Braun, who had been with Wire Card since 2000, became Chief Executive Officer (CEO).

In 2005, Wirecard was listed on the stock exchange via a reverse IPO with the listed InfoGenie AG, and the name changed to Wirecard. In a reverse IPO, an unlisted company is contributed to a listed company with no operating business. The unlisted company thus takes over the stock market listing (Peemöller et al., 2020).

This reverse IPO enabled Wirecard to avoid having its balance sheets audited, which would have been necessary in a normal IPO (Malcher et al., 2020). After Xcom Bank was also acquired in 2005, Wirecard AG received a banking license (Peitsmeier, 2018). This later became Wirecard Bank. Since 2018, Wirecard was also part of the German Share Index (DAX). After it became known that 1,9 billion euros were missing from the balance sheet, Wirecard had to file for insolvency in June 2020 (Grül et al., 2020).

Wirecard AG was divided into seven divisions in 2020. Most of these divisions included several subsidiaries both inside and outside Germany (Wirecard, 2020). Of particular importance for the accounting scandal were the subsidiaries Wirecard Technologies in Aschheim, Wirecard UK and Ireland Ltd. in Dublin and CardSystems Middle

East FZ LLC from Dubai (Lenz, 2020). The latter was headed by Oliver Bellenhaus. Bellenhaus was a responsible for Wirecard's third-party partner business, especially for the partner company Al Alam Solutions (Al Alam) (Bartz et al., 2020). On Wirecard's management board, Chief Operating Officer (COO) Jan Marsalek was primarily responsible for the third-party partner business (Bartz et al., 2020).

Due to the lack of banking licenses, Wirecard was dependent on third-party partners outside Europe. In addition to Al Alam, the most important of these third-party partners were the Philippine company PayEasy Solutions and the Singaporean company Senjo Payment Asia. Some of these third-party partners were also managed by former Wirecard employees (Grüll et al., 2020). A large part of the Wirecard Group's alleged profit was generated by this third-party partner business. However, these revenues largely did not exist. The existing part consisted largely of high-risk payments, primarily from pornography and gambling websites.

Incidents before 2020

Wirecard's annual financial statements were criticized several times even before 2020 (Bartz et al., 2020; Pemöller et al., 2020; Velte et al., 2021).

This influenced the stock market price, although those responsible for Wirecard itself were not prosecuted for a long time. Instead, critics were even accused of market manipulation. The influences on the stock market price are visualized in the appendix in Figures 1 to 5.

In 2008, allegations were made by the "Schutzgemeinschaft der Kapitalanleger" (SdK), a German shareholders' association.

Based on an analysis on the internet forum Wallstreet-Online, this association accused Wirecard of, among other things, misleading annual financial statements. This led to a sharp fall in Wirecard's share price, which lost more than half of its value. A short time later, it became known that representatives of SdK profited from this fall in the share price due to short selling (Peemöller et al., 2020). Ernst & Young (EY) was commissioned to conduct a special audit following the allegation. According to Wirecard, no material errors were found, but the report was not published. Following the special audit, EY became Wirecard's auditor (Drescher & Kirchner, 2008).

Another significant event followed in April 2015. The Financial Times (FT) journalist Dan McCrum published an article criticizing Wirecard for irregularities in its payment

behavior and balance sheets. Among other things, excessive advance payments were objected to in the acquisition of Asian companies. In the case of Trans Infotech, these amounted to over 80 % of the final price, which was significantly higher than the 5 % to 20 % used in some cases. In addition to the advance payments made, the high proportion of customer relationships in intangible assets was also viewed critically. Furthermore, differences were uncovered in the published consolidated balance sheets and in the corresponding balance sheets of the subsidiaries. Finally, discrepancies between the Group's balance sheet and cash flows were also criticized by the FT (McCrum, 2015). Following the publication of this criticism on April 27, 2015, the share price fell from over €40 to below €36. The share price fell by more than 10 %.

A report published by the analysis firm "Zatarra Research and Investigations" (Zatarra) in February 2016 also had a major impact, citing, among other things, allegations of money laundering and involvement in illegal gambling at Wirecard (Bartz et al., 2020). As a result of this report, as well as the FT's dissemination of the information, the share price fell by almost 15 % within one week, from over 42 euros to just 36 euros.

In December 2018, a criminal warrant was sought against the editor of the Zatarra report, Fraser Perring, for market manipulation. However, in May 2020, the court case was dropped for a fine after investigators found it difficult to prove criminal activity by Perring and others involved in Zatarra. Perring had denied the allegations of market manipulation.

At the beginning of 2019, the FT published information about an investigation launched by Wirecard into Wirecard's Singapore branch.

In this branch, falsified invoices were deliberately created in order to simulate non-existent cash flows to auditors. In addition, relationships with customers and suppliers were also faked. The person responsible for these falsifications was the head of accounting and financial operations in Asia, Edo Kurniwan. In addition to Kurniwan, however, Wirecard's board members were also involved, at least in part (McCrum and Palma, 2019).

Originally, the accounting and billing frauds were uncovered by a whistleblower, the head of Wirecard's legal department in Asia, Pavandeep Gill. He first informed the company's top management and, due to the latter's insufficient response, finally the FT

(Giesen et al., 2021). Following the FT's disclosure, Wirecard's share price plummeted by over 35 % in less than two weeks.

After commissioned investigators ruled out major accounting and billing fraud, FT journalists involved in the publication were reported by the German Federal Financial Supervisory Authority (BaFin) (Pemöller et al., 2020; Velte et al., 2021).

Significant for the disclosure of the accounting scandal was also, above all, an article by the FT from October 2019, where the newspaper harbored serious suspicions against Wirecard's partner company Al Alam in Dubai. Al Alam allegedly transmitted hundreds of millions of euros in payments each month and was responsible for half of Wirecard's profits in 2016. In reality, many of its customers either did not exist or had no business relationship with Al Alam or Wirecard. Moreover, Al Alam had hardly any employees (McCrum, 2019).

Wirecard managed Al Alam's business through its subsidiaries CardSystems Middle East in Dubai, and Wirecard UK & Ireland in Dublin, both of which also had hardly any employees (McCrum, 2019). Following the publication of the article, Wirecard's share price fell from 140 euros to 114 euros, i.e. by over 18 %.

After the article was published, Wirecard, at the urging of investor Softbank, commissioned the auditor KPMG to conduct a special audit (Bartz et al., 2020).

In December 2019, the FT published an article criticizing Wirecard for adding money in escrow accounts to its payments. This raised questions regarding the transparency and integrity of Wirecard's financial statements (McCrum, 2019).

KPMG special report

The special audit of Wirecard by KPMG started with the confirmation of the order at the end of October 2019. The aim of the audit was to investigate the allegations, particularly those published by the FT. It is also interesting to investigate whether revenue was knowingly falsely increased through fictitious customer relationships. In addition, the irregularities in Singapore and the inflated corporate purchase of an Indian company were to be investigated. During the investigation, the assignment was also extended so that the amount and existence of sales from third-party business were also investigated (KPMG, 2020).

For the investigation, KPMG inspected the documents of Wirecard's auditor, among other things. In addition, business partners in Dubai and the Philippines were interviewed (KPMG, 2020).

The results of the special audit were presented at the end of April 2020. During the investigation, the existence of the customers could be proven in most cases. According to the abbreviated designations originally published by the FT, some customers did indeed cease to exist from 2017 at the latest. However, according to Wirecard, these were assigned to companies that still existed at the respective times. However, these assignments were not always clear; in some cases, companies were assigned several different short designations. Wirecard also neglected to monitor customer compliance checks (KPMG, 2020). Based on current knowledge that large parts of Wirecard's Asian business probably did not exist, it can be suspected that the information handed over to KPMG was falsified and that controls may have been deliberately neglected. KPMG also complained that risks were not sufficiently identifiable in the management report in some cases, and that the sales in the trust accounts should not be recognized as cash or cash equivalents (KPMG, 2020).

According to an expert opinion, Wirecard's accounting was correct due to scope for interpretation and discretion under the International Financial Reporting Standards (IFRS), but the opinion made different assumptions to KPMG and therefore reached a different conclusion (Hübner, 2021).

The suspicion arises that the purpose of reporting revenues as part of cash and cash equivalents may have been to distort the balance sheet and cash flow statement, as this allowed cash and cash equivalents to be reported at a much higher level (KPMG, 2020).

Regarding Wirecard Singapore, KPMG criticized the lack of an internal control system for material transactions. Among other things, weaknesses were discovered in contract management and contract control. In some cases, contracts existed without economic substance, although this had already been discovered previously by EY as part of the 2018 annual audit (KPMG, 2020).

When it purchased the Indian company Hermes I Tickets Private Ltd. (Hermes) for over 200 million euros, Wirecard was criticized for the fact that Hermes had been sold

only one year earlier for 37 million euros to the fund, which then sold Hermes to Wirecard. KPMG was unable to identify the profiteer of this price difference. It also could not be ruled out that the selling fund was only acting as an intermediary (Boyd, 2018). This raises the suspicion that the company purchase served money laundering purposes.

KPMG is unable to confirm the existence of third-party partner revenue from 2016 to 2018. The culprits, according to the report, were deficiencies in Wirecard's internal organization, as well as the unwillingness of third-party partners to fully cooperate with the investigation. Transaction data, contracts between third-party partners and merchants, as well as account statements and bank confirmations often could not be provided to KPMG. The evidence provided in this regard was assessed by KPMG as insufficient for the investigation. Payments into escrow accounts worth around one billion euros could not be confirmed in this way (KPMG, 2020). From today's perspective, it is certain that these escrow accounts did not exist (Davies, 2020).

Following publication of the report on the special investigation, Wirecard's share price plummeted by over 30 %. Later, it also became known that KPMG had almost terminated the special investigation. This would have been catastrophic for Wirecard (Storbeck, 2020).

Only a few months after the publication of KPMG's report, EY had doubts about the fiduciary accounts, which ultimately led to the revelation of the fraud scandal (Bartz et al., 2020).

2.3 Linked data and knowledge graphs

Although the theoretical concept of knowledge graphs has been known before, their practical application gained importance through the introduction of Google’s knowledge graph in 2012. With this rising relevance there was also an increase in a variety of publications in the field. The result is a diverse collection of definitions and descriptions. Paulheim also recognized this in 2016 and in his study on knowledge graph refinement developed a set of characteristics that distinguish a knowledge graph from other data collections. "A knowledge graph (1.) mainly describes real world entities and their interrelations, organized in a graph, (2.) defines possible classes and relations of entities in a schema, (3.) allows for potentially interrelating arbitrary entities with each other and (4.) covers various topical domains.” (Paulheim, 2016).

As described by Kejriwal (2019), a knowledge graph can be described as a "graph-theoretic representation of human knowledge such that it can be ingested with semantics by a machine".

These semantics build the foundation of a knowledge graph and are expressed through triples. A triple allows to express relationships between two objects through three elements: (1) subject, (2) predicate, and (3) object. Figure 1 shows an example of a simple statement in triple form. First, the statement is expressed as a sentence. Second, the statement is shown in triple form. Third, the statement is presented in knowledge graph form.

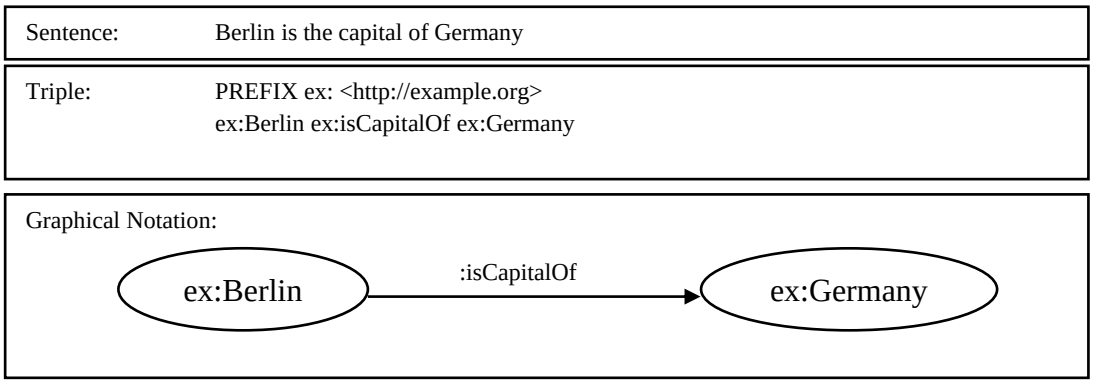


Figure 1: Simple statement expressed in triple form

As can be seen in Figure 1, the relationship between the two entities (Berlin and Germany) is expressed through a property *(isCapitalOf)*. Berlin is the subject of the triple, isCapitalOf is the predicate, and Germany serves as object. Berlin and Germany are resources and represented through a circle, while the property is depicted as a directed edge. The prefix 'ex:' abbreviates the URI 'http://example.org' and ensures that the triple adheres to the four principles of linked data, as proposed by Tim Berners-Lee (2009):

1. Use URIs to name things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things

However, the real world is often more complex than a simple statement or binary relation. In order to express n-ary relations, so called blank nodes are required.

As can be seen in Example 2 in Figure 2, blank nodes represent resources that do not have a URI or literals themselves and can only be used as subject or object in a triple. They can be either declared with an underscored prefix (_:example) or abbreviated through squared brackets. Example 2 also contains so called literals, which represent values. According to RDF standards, literals can only be used at the object position.

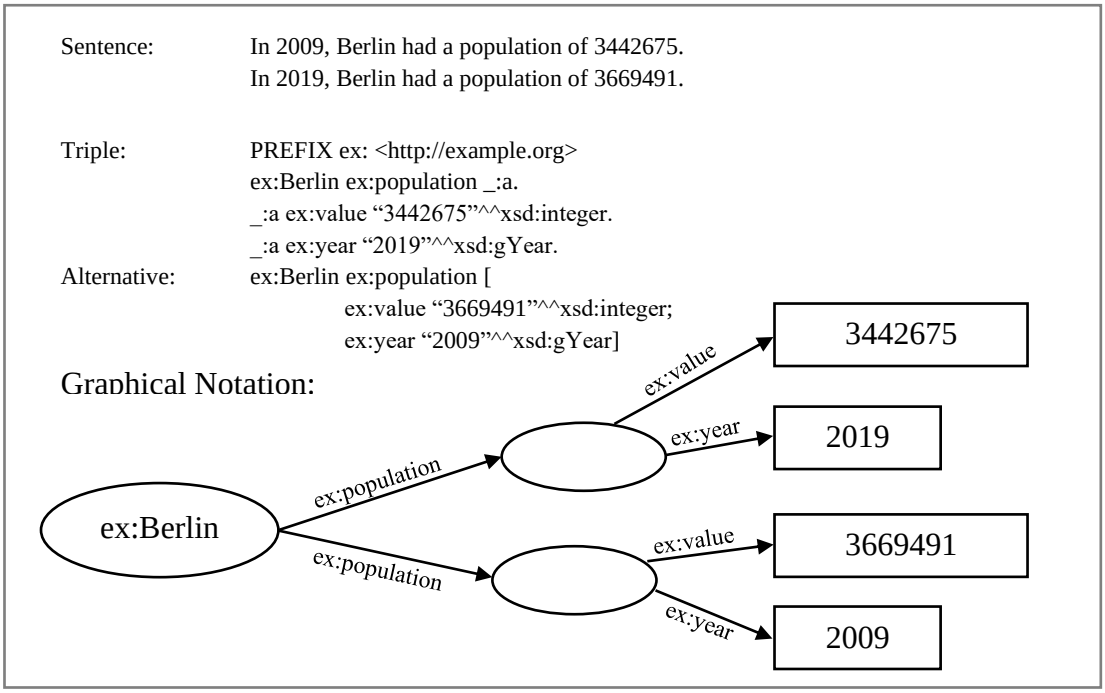


Figure 2: n-ary statement expressed using blank nodes

In order to query information from an RDF triple construct, the "SPARQL Protocol and RDF Query Language" (SPARQL) is utilized. In general, there are four types of queries:

Query type	Description
ASK	Return true or false, depending on whether there is a graph pattern
SELECT	Return all or a subset of the solution mappings
DESCRIBE	Return a set of triples / a graph that describes a certain resource (URI)
CONSTRUCT	Return a set of triples/a graph, where the mappings are filled into a specific graph pattern template

Table 1: Main types of SPARQL queries

3 Data and research methodology

The following section provides an overview of the research methodology and its objectives. The systematic approach followed to build a knowledge graph and complete the paper can be divided into methodical steps. These steps were adopted after review of current literature together with best practices in field.

The overall objective of this paper is to explore the use of knowledge graphs as a tool to investigate the people behind the Wirecard scandal. It is further interesting to examine how diverse the members of the supervisory board and the management board are in terms of criteria such as place of birth, education and professional experience.

It may be possible at the end to understand whether there are differences in terms of the composition and members of the supervisory board and the management board that allow further conclusions. The results are also compared with the members of the supervisory and management boards of the other DAX30 companies. As such, there are five subgoals that also serve as milestones: (1) database generation (2) graph modeling (3) graph generation (4) graph analysis, and (5) graph visualization.

The section *Data* includes extracting new data from various sources and creating a central database that adheres to common database normalization rules and serves as the foundation of the knowledge graph. *Graph modeling* comprises the design of a model based on the available data that allows for meaningful graph traversal and analysis. *Graph generation* describes an instantiation of the graph model. *Graph analysis* describes all tasks related to finding meaningful insights and evaluating the paper itself.

Graph visualization aims to provide some visual representation of the graph that allows for interaction and navigation. As such, the goal is to use programs that provide a Graphical User Interface (GUI) and therefore remove entry barriers for users with less technical expertise.

3.1 Data

Working with data always includes some difficulties and requires special care. This is even more important to implement in the paper mentioned here, as it concerns personal data. For this reason, special care has been taken not to violate the privacy of the persons in the database. The first part of the chapter is dedicated to the origin of data.

Then the scope of the data is discussed. The following parts deal with the storage and linking of the data.

3.1.1 Data origin

For each person, a variety of data was collected, such as year and place of birth, current position, secondary employment as well as the last three professional positions. In each case, the data is based on the data published on the firm website. Some companies offer a very detailed CV in PDF format. Other companies, however, provide only a short text as a brief description. In those cases, additional information was hand-collected from other sources, for instance professional social networks like LinkedIn.

It was therefore challenging to scrape data or semi-automate the process. This made data collection very time-consuming and tedious. In addition, it is always important to pay attention to data quality when collecting data manually. Since the data was collected by different people, discrepancies can also arise easily. We tried to leverage open graph databases such as DBpedia[3] and WikiData[4] but found the information on persons of interest rather sparse and outdated. Furthermore, data was sourced from Bloomberg. Additionally, further information on the members of the supervisory board and the management board derives from the internet. We assumed that the information on firm websites is correct. However, this is not always the case for the information we found on the internet. Consequently, we verified information from non-official sources.

3.1.2 Data scope

We base our examination of German companies listed in the DAX30 in the year 2019. Only those members of the supervisory board or management board who still held their office at the end of the 2019 financial year are considered. Changes during the year (e.g. board member retirements) are not taken into account.

Eventually, we collect and store data of 1.173 persons of interest in our database, including supervisory board members, management board members and auditors. Using OpenRefine[5] we were able to link about a third (32 %) to Wikidata.

[3] https://wiki.dbpedia.org/.
[4] https://www.wikidata.org/wiki/Wikidata:Main_Page.
[5] https://openrefine.org/

Furthermore, we captured 1.173 companies in the database and were able to link more than half (53 %) to Wikidata. In total, our database includes 5.116 positions with 1.830 held as of December 2020 as well as 1.128 educational degrees or programs at 341 universities completed by persons of interest. Data provided by the German financial supervision agency BaFin allowed capturing 354 dealings made by persons of interest between December 2019 and December 2020. Figure 3 gives an overview of the data.

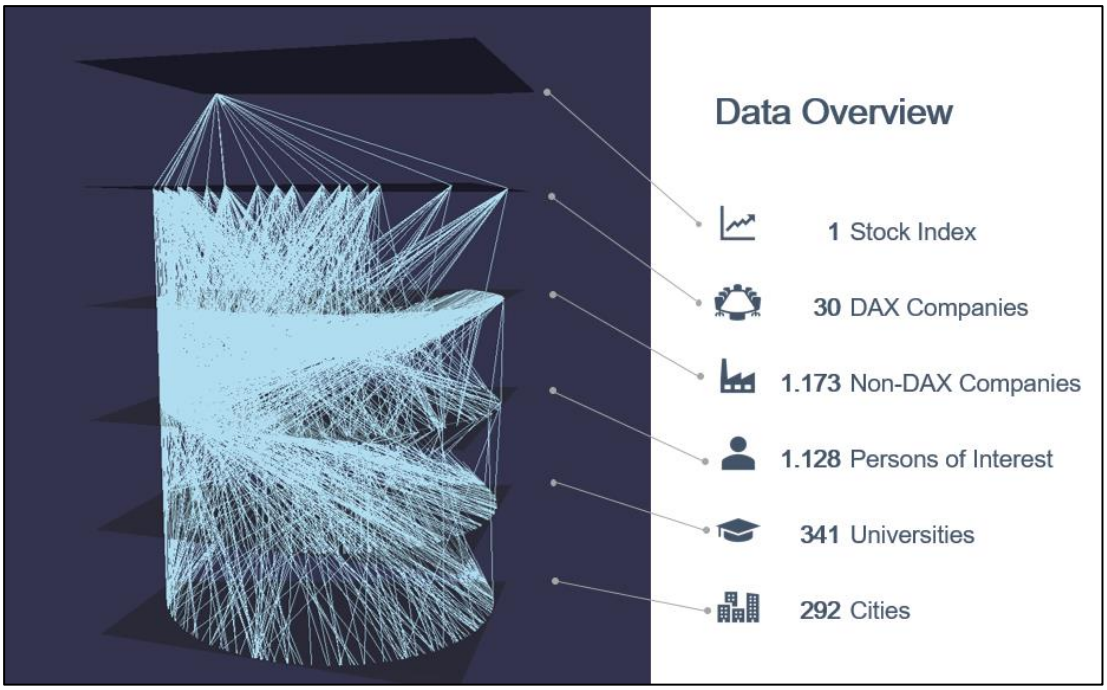


Figure 3: Data overview

3.1.3 Data processing

Especially the work with data was not easy in this case. On the one hand, there is a large amount of different data per person, but on the other hand, there is also a large number of people in the database. On top of that, the data often comes from different sources and has different data types. In order to link the different data types with each other, they had to be normalized. For instance, different main categories for degrees have been created. In Germany, for example, there are various degrees that can be grouped under the main category "Business". These include, for example, the degree “Volkswirtschaftslehre”, “Betriebswirtschaftslehre”, but also the English term “Business Administration”. Table 2 gives an overview of the grouped study programs.

Field 1/ Field 3	Field 2/ Field 4
Wirtschaftswissenschaften	VWL BWL FACT Business Administration Kommunikationswissenschaft Marketing Economics
Ingenieurswissenschaften	Elektrotechnik Informatik Maschinenbau Maschinenbau Luft- und Raumfahrttechnik Agrartechnik
Naturwissenschaften	Biologie Chemie Physik Mathematik Medizin
Geisteswissenschaften	Theologie Philosophie Journalismus Politikwissenschaften Soziologie Geschichte Psychologie Anglistik Kunst Sport
Rechtswissenschaften	
Pädagogik	
Linguistik	Englisch, etc.
Other	

Table 2: Overview of education categories

The example of degree programs illustrates the underlying principle of normalizing the data in this paper. However, this was not only important for study programs, but for each of the categories in Table 3. Table 3 presents and explains the different categories of the database.

Category	Description
Person	Assigns a unique ID to each person *(personEntity)*. Contains personal information of each person in the graph (e.g., name, year of birth).
Company	Assigns a unique ID to each company *(companyEntity)*. Contains information of each company in the graph (e.g., name, ISIN, DAX Membership).
University	Assigns a unique ID to each university *(universityEntity)*. Contains information of each university in the graph (e.g., name, link to other databases).
Cities	Assigns a unique ID to each city *(cityEntity)*. Contains information of each city in the graph (e.g., name, link to other databases).
Countries	Assigns a unique ID to each country *(countryEntity)*. Contains information of each country in the graph (e.g., name, link to other databases).
Position	Contains information on employee positions a person has had or currently has in their career with a company. Serves as link between a company and a person.
Education	Contains information on educational programs or degrees a person has had. Serves as link between a university and a person.
Audit	Contains information on audits a company has had. Serves as link between an auditing company, the company being audited and the auditors (a person).
Dealings	Contains information on dealings (German: Eigengeschäfte) a person has had with a company. Serves as link between a company and a person.
Relationships	Contains information on relationships between persons (e.g., spouse, sibling). Serves as link between two persons.
Organization	*Discontinued due to sparse data: Contains information on a person's membership in organizations. Serves as link between a person and an organization.*

Table 3: Overview of information in the database

The categories in Table 3 also form the basis for the structure of the database. All tabs related to resources (highlighted in Table 3 in blue) are assigned a unique ID that is converted to a Uniform Resource Identifier (URI) in the graph generation as well as links to external databases (e.g., *wikidata*_sameAs) if applicable.

The *position* tab contains the column '*divisionName'* that allows to specify the suborganization of a position within the company. Furthermore, it has three Boolean indicators (*current*, *positionAR*, *positionVO*). The column *current* indicates whether a role is currently held (1 = true, 0 = false). Consequently, positions currently hold do not have an end date. Furthermore, this allows us to indicate that a position is not current, even when the exact end date is not known.

The column *positionAR* indicates whether a position is related to some form of supervisory board (1 = true, 0 = false). The column *positionVO* indicates whether a position is related to some form of management board (1 = true, 0 = false).

3.1.4 Data linking

The data from the database should also be linked to external data sources. This increases the scope of data, data quality and simplifies data research. While linking entities in the graph to databases such as WikiData or DBpedia can be done manually, a useful tool to semi-automate this step is OpenRefine. OpenRefine matches strings of names in the file to names of entities in a selected database. Here it is possible to choose to auto-match candidates with high confidence.

3.2 Research methodology

Graph modeling

For the creation of the knowledge graph, the ontology is an important aspect. Ontologies in computer science are usually linguistic and formally ordered representations of a set of concepts and the relationships that exist between them in a particular subject domain. They are used to exchange knowledge in a digitized and formal form between application programs and services (Robinson et al., 2013). The graph model is depicted in Figure 4 in the Appendix.

The ontology has been aligned with already existing ontologies. The goal of the ontology used is to accommodate as much information as possible about the boards of directors and supervisory boards in an orderly fashion. A few of the categories mentioned above require a precise arrangement of the data. For example, the different professional positions are difficult to capture, as they sometimes appear very similar but then show differences. Special care must be taken that the data follows a clear scheme and that there are no ambiguous entries. In addition, duplicate entries should be avoided. However, the graph model created in Figure 4 provides a concise basis for information.

Graph analysis

Knowledge graph analysis offers several advantages over other databases. In our case, for example, Excel makes it difficult to find or analyze links between boards of directors and supervisory boards. This is due to the fact that the information in the rows and columns is difficult to model and, in particular, it is difficult to identify connections. With the help of the knowledge graph, this is not a problem. There are two different approaches to analyze the knowledge graph.

First, queries are performed to analyze the persons of interest. In order to execute SPARQL queries on the graph, we can use GraphDB[6] or Blazegraph[7]. Especially the SPARQL queries offer an added value to other databases. Here it is possible, for example, to identify with one query who has worked together in a company.
Second, visualizations help to get a better understanding of special cases. The different possibilities of visualization will be presented in the following.

In addition, however, the data can be analyzed otherwise. For example, the question arose where the members of the management and the supervisory board come from. In particular, birthPlaces_map may illustrate the use of open graph databases beyond our own data, as it queries city coordinates from Wikidata and therefore allows us to map birthplaces of DAX members.

Figure 5 shows the map created using Tableau with the birthplaces of the members of the management and supervisory board of the DAX30 companies.

[6] https://www.ontotext.com/products/graphdb/.
[7] https://blazegraph.com/.

Figure 5: Map depicting birthplaces of DAX management board members (blue) and supervisory board members (red)

Graph visualization

Visualization is not the key part of knowledge graphs. However, in our analysis we visualized connections for better understanding. Especially when looking at the connections of individuals, this can be very helpful. Different tools offer different ways of visualization. In the following, we present several tools.

Gephi

Gephi[8] is an intuitive and established tool for visualizing graphs and offers a variety of different network analysis functions. On the downside, the tool does not scale very well and has difficulty displaying large graphs. First, the turtle graph (or subgraph) has to be converted in .gexf format.

[8] https://gephi.org/.

The scripts with the extension _to_gexf in execgraph/python_conversion show this procedure. In Gephi the .gexf files can be imported and displayed in various forms (e.g., graph visualization, data table) and personalized (e.g., color sche me, node size). Furthermore, different network analysis metrics can be calculated, with the most useful one probably being PageRank. The PageRank metric proposed by Page et al. (1999) calculates the perceived importance of a node in a directed graph and can be expressed by the following simplified algorithm:

$$R(u) = c \sum_{u \in B_u} \frac{R(v)}{N_v}$$

With u being a node in the graph, v being a node in the set B_u containing all nodes linking to page u, and N_v being the number of links from node v. The PageRank of node u is then iteratively calculated by summing up the PageRank of each node linking to node u divided by their outgoing links. As such, links of nodes with higher PageRank are weighted more than links of those with lower PageRank. The initial PageRank value for each node in Gephi is set to $\frac{1}{n}$, with n being the number of nodes in the graph. Figure 6 shows the visualization of the members of the supervisory board and the management board of the DAX30 companies with Gephi.

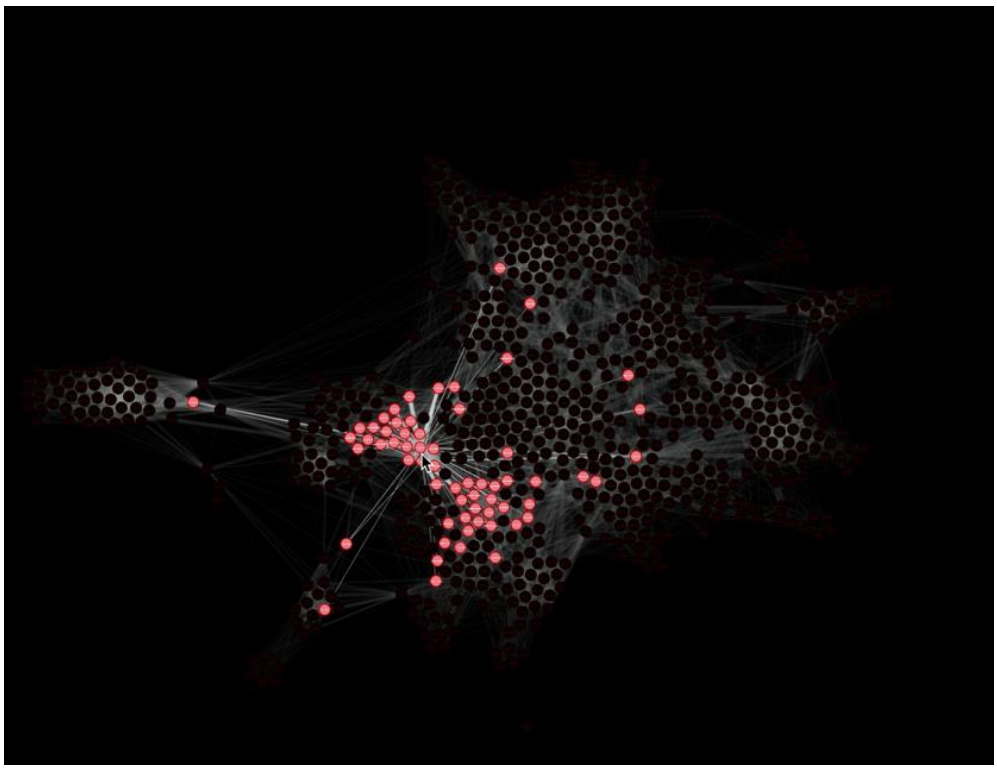


Figure 6: Screenshot of Gephi visualization

Graphistry

Graphistry[9] is a more advanced tool for visualizing graphs that can be used in Python, which has the best performance of the tools mentioned in this section. On the downside, the tool does not have as many network analyses functions as Gephi. Figure 7 gives an example of how Graphistry visualizations look like. It shows the 1st degree connections between people in the database. The visualization of Graphistry on the right side in Figure 7 shows that by clicking on a person the graph changes and visualizes only the relevant information.

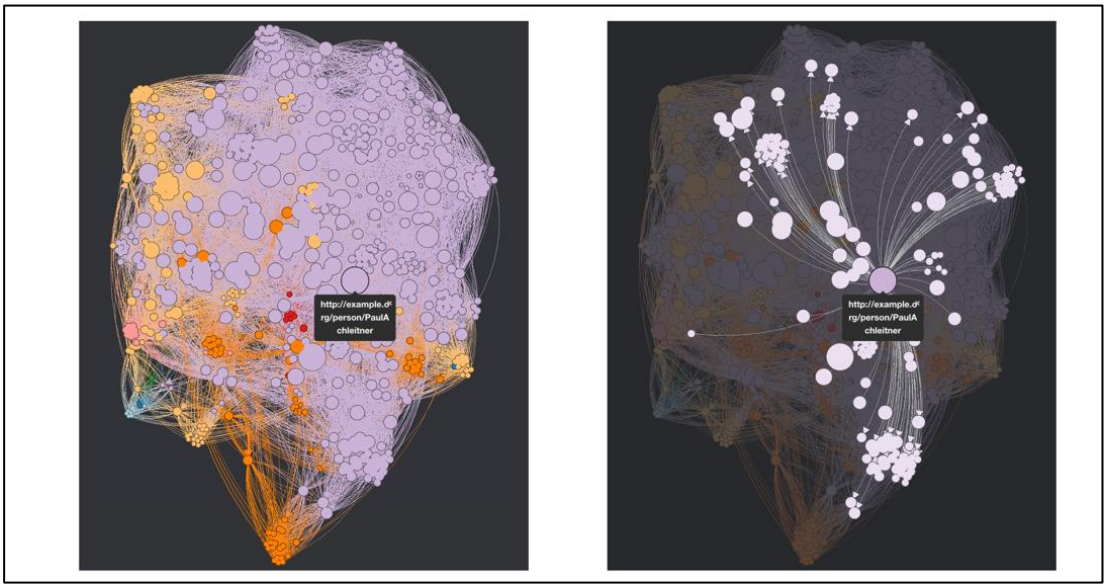


Figure 7: Screenshot of Graphistry visualization

Tarsier

Tarsier[10] is a tool developed for 3D RDF visualization as it allows arranging data on multiple levels. Visualization with different levels is well suited for our work. This makes it possible for the members of the supervisory board, members of the management board and potential contacts to be represented at different levels. In this way, the visualization appears clearer and the various connections can be better represented. However, before using tarsier a more advanced setup is required. Python, Java, and Blazegraph[11] are required for this tool. Figure 8 shows how Tarsier is used to display the entire Wirecard database in a graph.

[9] https://www.graphistry.com/.
[10] https://github.com/desmovalvo/tarsier.
[11] https://blazegraph.com/.

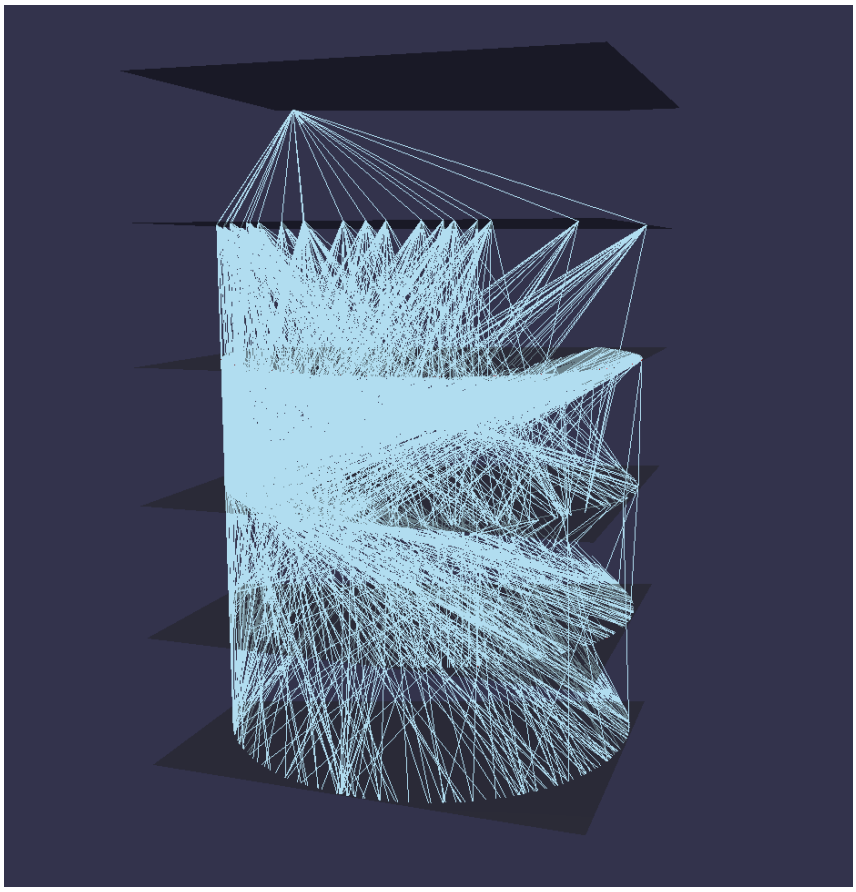


Figure 8: Screenshot of Tarsier visualization

4 Analysis

4.1 Wirecard statistics

Our graph contains 10 persons of interest related to Wirecard. These persons of interest include management board members, supervisory board members and auditors. 32 % of these people have a link to Wikidata and 39 % have a personal LinkedIn page.

In terms of education, for the people related to Wirecard we have identified 18 degrees or educational programs at 15 different universities.

For the persons of interest, we count 94 roles with division, title, start and end year. In total there are 1.830 currently held roles in our knowledge graph, including 510 management roles and 1.071 supervisory board roles. The knowledge graph shows 215 potential connections to other people based on previous roles, education and place of birth.

4.1.1 Supervisory board

In the case of a scandal like this, the question arises whether the monitoring bodies have performed their duties diligently. This also includes the supervisory board of Wirecard AG. The duty of care for supervisory board members is governed by Section 116 of the German Stock Corporation Act (GSCA) in conjunction with Section 93 GSCA. Based on this, it is questionable whether the members of the supervisory board of Wirecard AG have fulfilled "the due diligence of a prudent and conscientious manager" within the meaning of Section 93 GSCA. Diligent would also be the fulfillment of the duties and rights of the supervisory board pursuant to Section 111 GSCA. Tina Kleingarn left the supervisory board of Wirecard AG after just one year. She reported that it had been extremely difficult to inspect documents and generally perform her duties as a supervisory board member diligently. She saw deficiencies and potential risks due to poor corporate governance and warned the remaining members of the supervisory board of Wirecard AG in her resignation letter (Slodczyk, 2020). The question therefore arises as to why the other supervisory board members did not act similarly, as they were potentially hindered in their exercise as a proper and conscientious supervisory board member.

For this purpose, the composition of the supervisory board of Wirecard AG in recent years is examined in more detail on the basis of Wirecard's annual reports. Two facts are striking about the composition of the supervisory board. Firstly, the, comparatively, small number of members and secondly, the members themselves. According to DSW supervisory board studies from 2018 to 2020, the number of members in DAX 30 companies on the supervisory board averaged 15 to 16 members (DSW, 2018, 2019, 2020). In 2018, Wirecard AG acted with five members in the interim following the departure of Tina Kleingarn on December 31, 2017.

A lower number of supervisory board members does not equate to insufficient monitoring when looking at the companies in relative terms, but it would possibly be desirable to come closer to the average of the other DAX 30 companies. The DSW 2020 study also shows that newcomer Delivery Hero also has only 6 supervisory board members. Almost half of these people also have no experience whatsoever in supervisory board activities. The composition is like that of Wirecard AG.

Matthias Wulf, Alfons W. Henseler and Stefan Klestil formed the supervisory board since 2009. In 2019 Alfons W. Henseler stepped down and was replaced by Thomas Eichelmann. The new members Tina Kleingarn and Vuyiswa V. M'Cwabeni, who were appointed in 2016, had no experience in supervisory board activities at the time of their appointment. It is questionable whether the composition with three long-standing and two inexperienced supervisory board members was deliberately designed by the members of the management board to maintain control and avoid conflict. Pursuant to Section 95 (4) GSCA, the increase in the number of members was not limited by a lack of prerequisites. Instead, Wirecard AG always adhered to the minimum number of three supervisory board members pursuant to Section 95 (1) GSCA. Anastassia Lauterbach and Thomas Eichelmann were not appointed to the Wirecard supervisory board until 2018. These two persons have the most experience in supervisory activities. In retrospect, one could speak of an acquisition of well-known members for image reasons and not of a serious increase in monitoring.

4.1.2 Management board

In the German two-tier system, the management board is responsible for managing the company under its own responsibility (Section 76 (1) GSCA). Pursuant to Section 108 (1) GSCA, appointments are made by resolution of the supervisory board with a simple majority of votes. The maximum term of appointment is five years, whereby an extension or reappointment is permissible in accordance with Section 84 (1) GSCA for a further maximum of five years in each case. The purpose of this provision is to ensure that the supervisory board regularly assesses the quality and work of the management board in order to decide on its continuation in office. In the event of the appointment of more than one member of the management board, the supervisory board is authorized under Sec. 84 (2) GSCA to appoint one member as chairman of the management board.

The following section deals with the management board of Wirecard. The management board consists of 4 people:

- Dr. Markus Braun
- Susanne Steidl
- Alexander von Knoop
- Jan Marsalek.

Even before the Wirecard scandal, Wirecard had a special composition of the management board. Dr. Markus Braun and Jan Marsalek stood out in the process of data gathering. Limited information was available for these people. We gathered 48 triples of information for Dr. Markus Braun and 43 triples for Jan Marsalek. For comparison, for Oliver Bäte, CEO of Allianz SE, we gathered 123 triples of information. The lack of information for the people of Wirecard are not limited to a certain field. For instance, for Jan Marsalek, it is not possible to find any information about his education or the place of birth. In addition to that, it is also not possible to find pictures or videos from Jan Marsalek. Just one picture of him was available on the company website.

4.2 Connections

As already mentioned, our database contains all supervisory board members, management board members and auditors of the DAX30 companies. The aim of the following analysis is to show connections between these individuals. Various diagrams have been created for each of these connections. These help not only in analyzing the connections, but also in visualizing them. Visualizations provide a good way to better analyze and interpret connections. Several analyses and visualizations are presented below.

4.2.1 Industry connections

There are several ways in which people can be connected to each other. One possibility is industry connections. This means that all kind of connections between the management board, supervisory board and potential work and educational connections are visualized. Potential connections are connections where it is not certain if the people really know each other. For example, people may have worked in the same company at the same time or studied at the same university, but we cannot be sure whether they know each other. In the example we can see that there are different links.

However, we can also see that the links refer to specific people. Figure 9 shows the visualization of Wirecard's industry connections. The graphic, like the following ones, is visualized with Tarsier. This has the advantage that different layers can be used. The first level is the company Wirecard, followed by the levels for the management board and the supervisory board of Wirecard. This is followed by levels for potential work connections and educational connections.

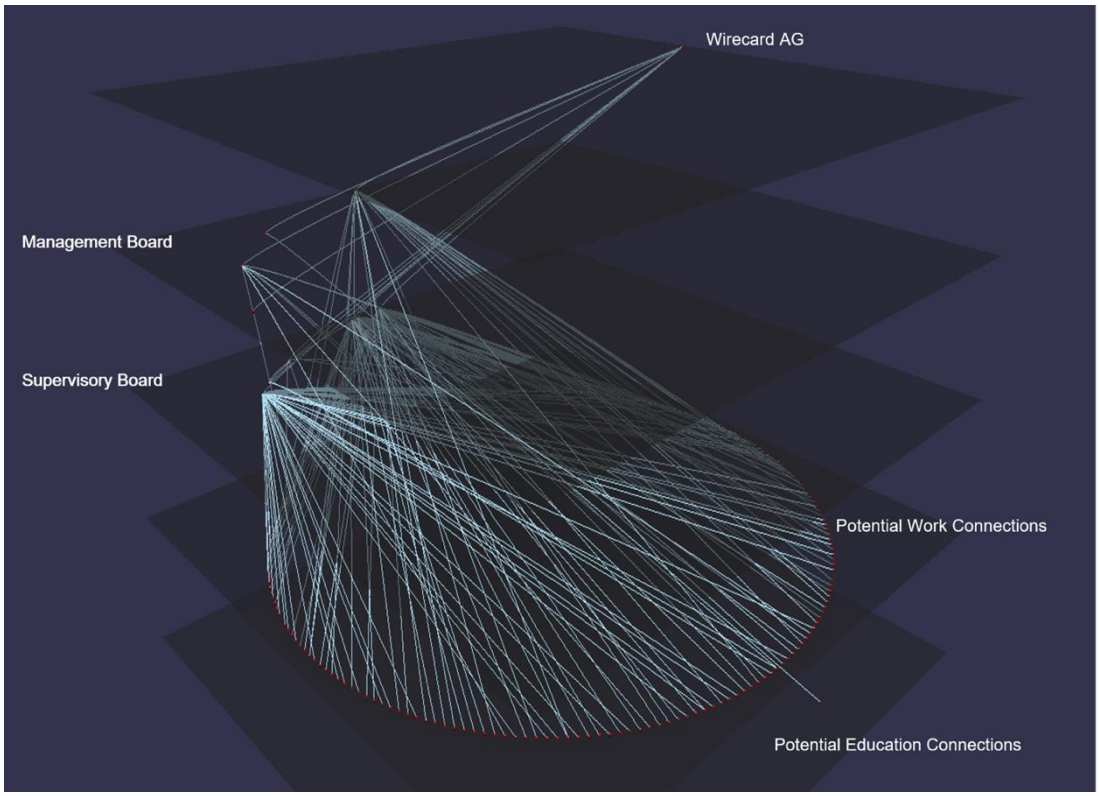


Figure 9: Visualization of Wirecard industry connections

The graph shows that there are a variety of different connections between the different levels. However, it can also be seen that most of the connections from the management board can be traced back to two people. The situation is similar on the supervisory board. It is also possible to identify a large number of potential connections between members of the supervisory board and the management board with other individuals from our database in terms of joint professional activity. Regarding joint education, only a few connections can be identified in Figure 9. To have a better context, we have done the same analysis with Allianz SE. Figure 10 shows the visualization of Allianz's industry connections. This is the same methodology as in Figure 9, but with Allianz as the company.

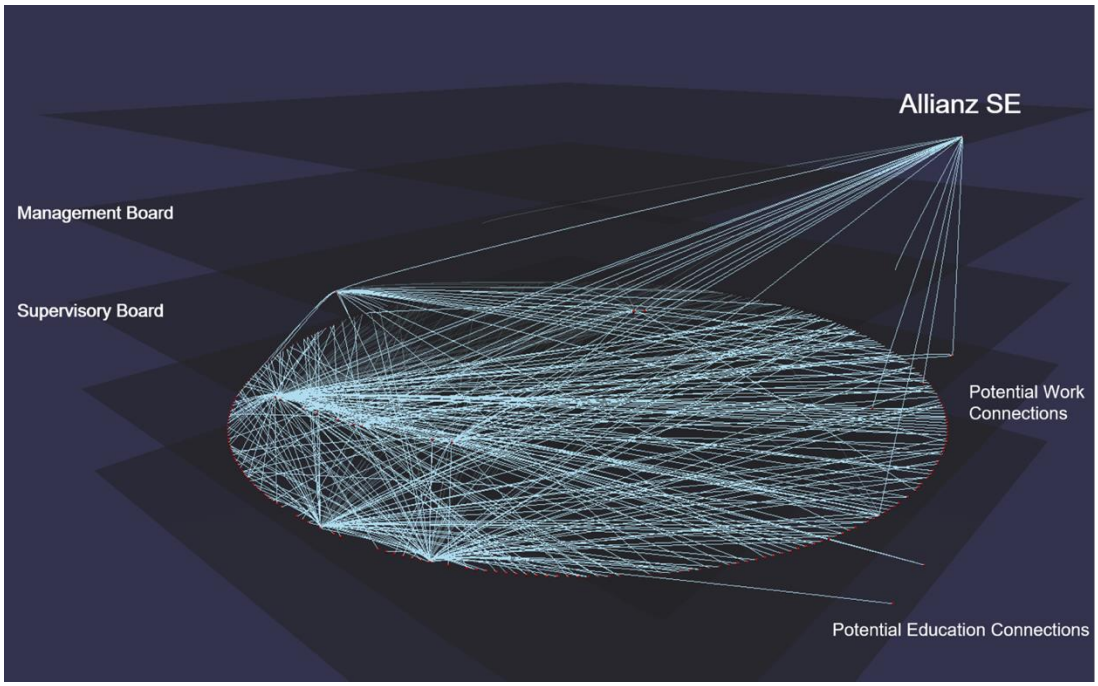


Figure 10: Visualization of Allianz industry connections

The visualization in Figure 10 shows that there are significantly more connections between the different levels at Allianz. However, there are only very few connections in terms of joint education here as well. Nevertheless, there are many connections between members of the supervisory board, the management board and potential professional contacts. This initial analysis provides a good overview. But it is necessary to take a closer look at the connections between the people in more detail. Therefor the next visualization shows the inter-board connections.

4.2.2 Inter-board connections

Inter-board connections show links between members of the supervisory board and the management board. These can be both direct and indirect. Direct connections are connections that exist between the supervisory board and the management board. Indirect connections are connections that potentially exist via third parties.

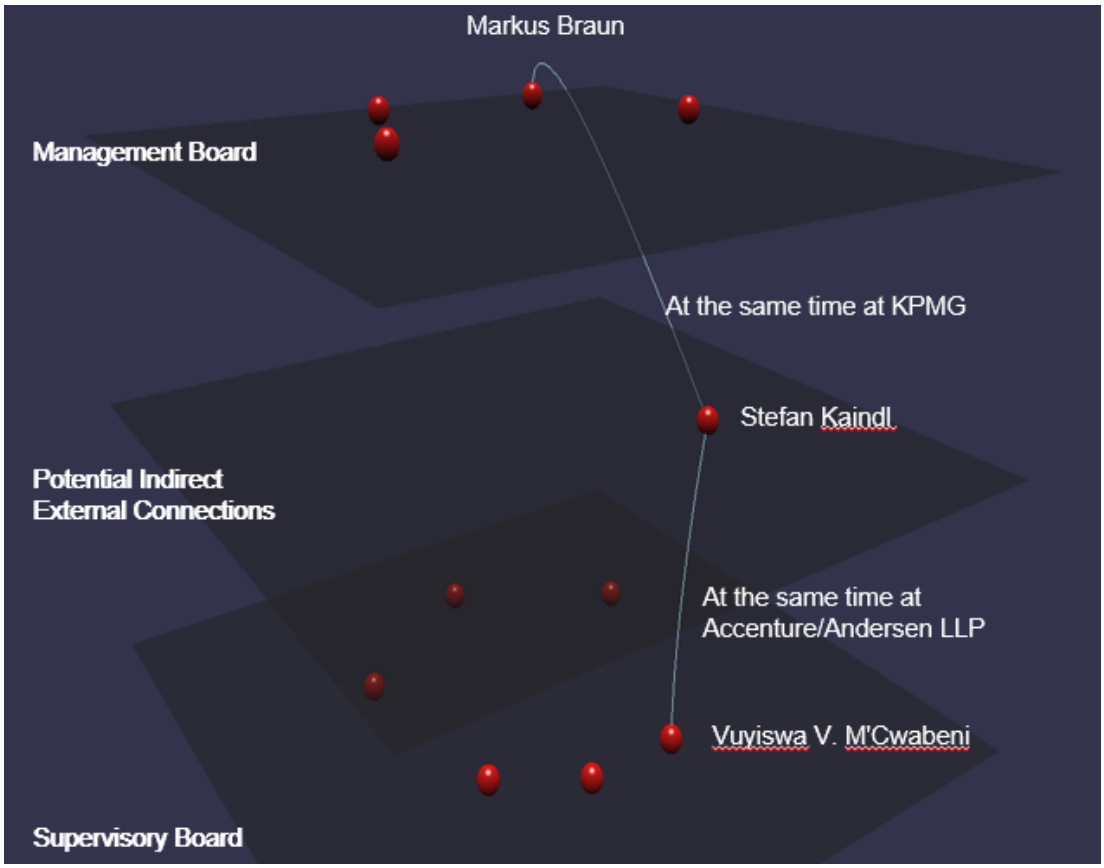


Figure 11: Visualization of Wirecard inter-board connections

In Figure 11, it is clear to see that Markus Braun is the only person from the management board who has contact with a person from the supervisory board via a third person. However, this link is not particularly meaningful. Stefan Kaindl worked at KPMG at the same time as Markus Braun, but they worked at different locations. The same applies to the link between Stefan Kaindl and Vuyiswa V. M'Cwabeni. Otherwise, there are no links between the persons. It is noticeable that there are also no direct links between the management board and the supervisory board. In the following, the results are again compared with the results of the Allianz.

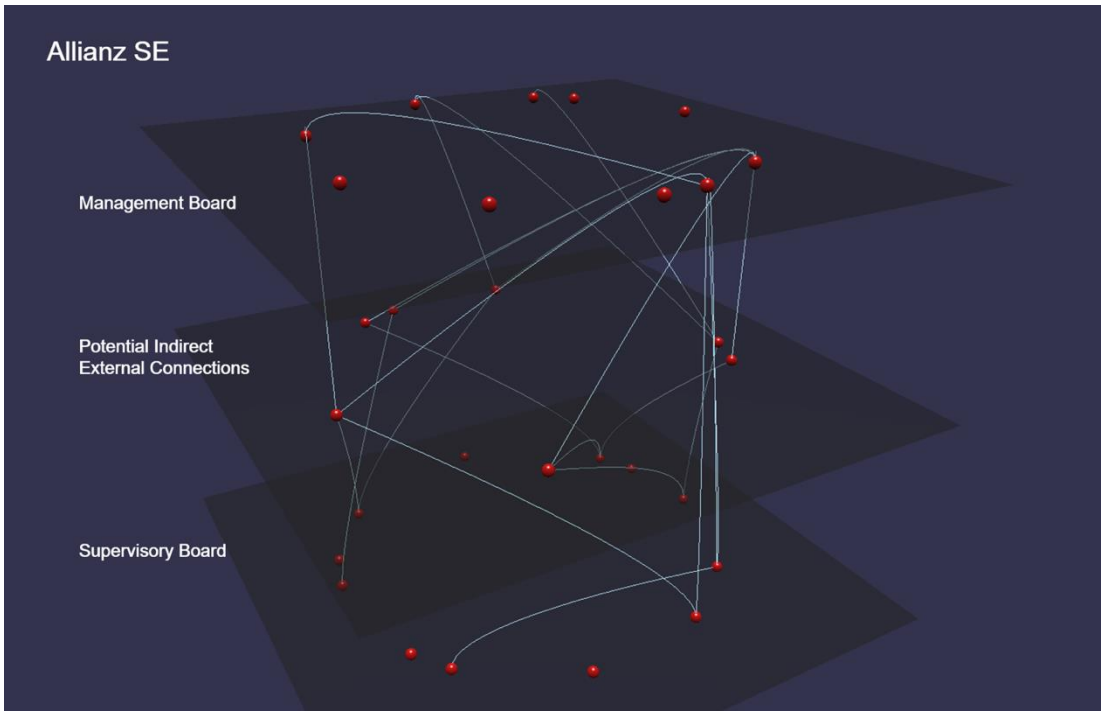


Figure 12: Visualization of Allianz inter-board connections

In this example, too, it is clear to see that there are significantly more links at Allianz. This is also not particularly unusual, as there is only one connection in the Wirecard example. In the Allianz example, on the other hand, many connections can be observed. These links exist in all directions. There are connections between the members of the management board, between members of the supervisory board, but also between the management board and the supervisory board. In addition, there are also potential indirect connections between the members of the supervisory board and the management board via third parties. It can also be seen that the connections are not limited to a few people but exist at all levels among different people.

In this context, it is questionable why there are so few connections at Wirecard. One reason for this can be observed in the following example in Figure 13. Figure 13 visualizes the experience in board activities of the members of Wirecard's supervisory board and management board.

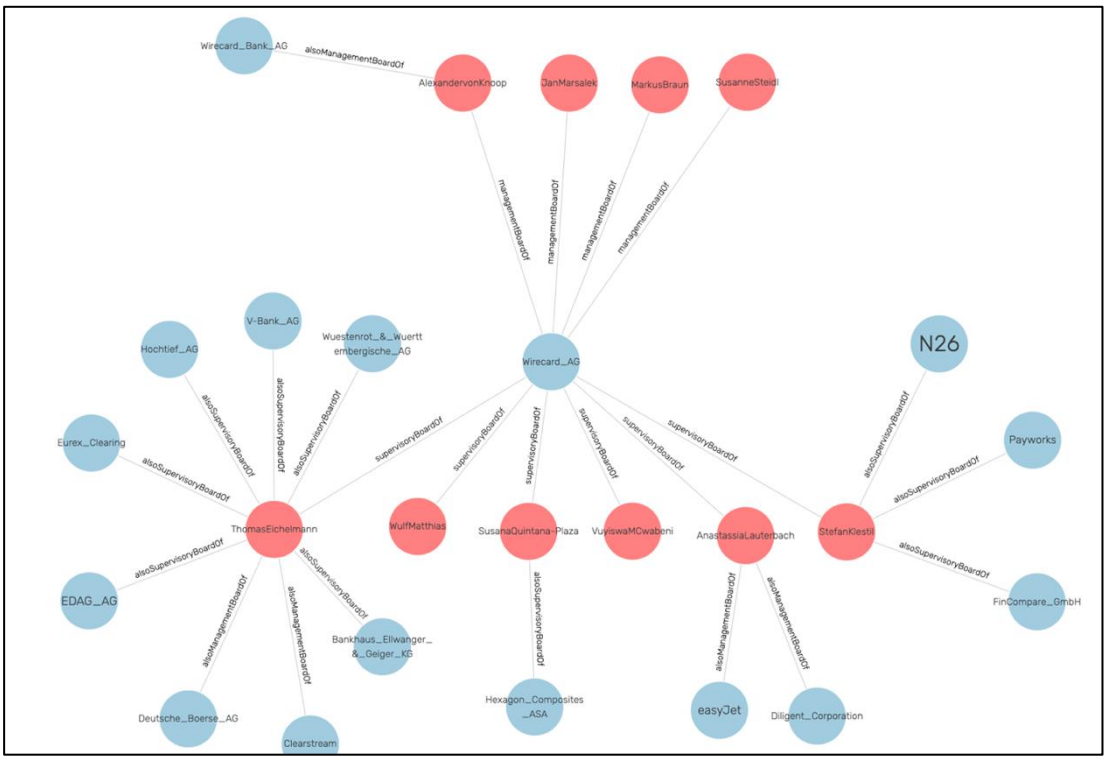


Figure 13: Visualization of the experience in board activities of the members of Wirecard's supervisory board and management board

The members of the board can be found in the upper part of the visualization. Of the four members, none has yet gained experience in another supervisory board or management board of a DAX30 company. Only Alexander von Knoop has experience on another board. Unfortunately, this was only at Wirecard Bank. Accordingly, it can be stated that none of the members of Wirecard's management board can demonstrate any notable experience in managing a company. This, of course, does not speak well for the company and could be a reason for the collapse.
However, it can already be found in the literature section that the management board of a German stock corporation is appointed by the supervisory board.

It is now questionable why the supervisory board appointed or extended the term of office of the management board member, even though he or she cannot demonstrate any significant experience. To answer this question, it is interesting to look at the experience of the members of the supervisory board.

The members of the supervisory board can be found in Figure 14 at the bottom. Two facts are notable about the composition of the supervisory board. Firstly, the comparatively small number of members and secondly, the members themselves.

According to the DSW supervisory board Studies from 2018 to 2020, the number of members in DAX 30 companies on the supervisory board averaged 15 to 16 members (DSW, 2018, 2019, 2020). In 2018, Wirecard AG acted with five members in the interim following the departure of Tina Kleingarn on 31 December 2017. A lower number of supervisory board members does not imply inadequate supervision when considering the companies in relative terms, but it might be preferable to bring the number closer to the average of the other DAX 30 companies. Table 4 in the Appendix shows the amount of the members of the supervisory board, members of the management board and auditors of the DAX30-Companies.

Moreover, almost half of these people have no experience in supervisory board activities. The composition is similar to that of Wirecard AG. But before taking a closer look at the experience of the supervisory board, it is interesting to look at the composition of the supervisory board over the years. Figure 14 provides an overview of the composition of the supervisory board from 2013 to 2019.

Figure 14: Composition of the Wirecard supervisory board from 2013 – 2019

Figure 14 shows that Matthias Wulf, Alfons W. Henseler and Stefan Klestil have formed the supervisory board since 2009. It was not until 2019 that Alfons W. Henseler stepped down and was replaced by Thomas Eichelmann. The new members Tina Kleingarn and Vuyiswa V. M'Cwabeni, who were appointed in 2016, had no experience in supervisory board activities at the time of their appointment. Matthias Wulf also had no experience in board activities. Stefan Klestil already had some experience in board activities.

He did not gain this experience in DAX30 companies, but in the fintech sector. This was of course advantageous for a company like Wirecard. Dr Anastassia Lauterbach also has quite a bit of experience. She has been a member of various management and supervisory boards. Most of these were not DAX30 companies, but she still had a lot more experience than the other members of the supervisory board. Stefan Eichelmann has already served in various board positions and in various supervisory boards. However, Stefan Eichelmann only joined Wirecard's supervisory board in 2019. He then also led the supervisory board as the chairman. However, it is questionable to what extent he should have prevented the collapse of the company.

Table 5 gives a further overview of the activity and duration of other board activities of the members of the board.

Person URI	Company URI	Mgmt Position	Sup Position	start	end	duration in years
http://execgraph.org/AnastassiaLauterbach	http://execgraph.org/Diligent_Corporation	1	0	2018	2020	2
http://execgraph.org/AnastassiaLauterbach	http://execgraph.org/easyJet	1	0	2019	2020	1
http://execgraph.org/StefanKlestil	http://execgraph.org/FinCompare_GmbH	0	1	2017	2020	3
http://execgraph.org/StefanKlestil	http://execgraph.org/N26	0	1	2013	2020	7
http://execgraph.org/StefanKlestil	http://execgraph.org/Payworks	0	1	2016	2020	4
http://execgraph.org/SusanaQuintana-Plaza	http://execgraph.org/Hexagon_Composites_ASA	0	1	NULL	2020	N/A
http://execgraph.org/ThomasEichelmann	http://execgraph.org/Bankhaus_Ellwanger_&_Geiger_KG	0	1	2012	2018	6
http://execgraph.org/ThomasEichelmann	http://execgraph.org/Clearstream	1	0	2007	2009	2
http://execgraph.org/ThomasEichelmann	http://execgraph.org/Deutsche_Boerse_AG	1	0	2007	2009	2
http://execgraph.org/ThomasEichelmann	http://execgraph.org/EDAG_AG	0	1	2010	2018	8
http://execgraph.org/ThomasEichelmann	http://execgraph.org/Eurex_Clearing	0	1	2007	2009	2
http://execgraph.org/ThomasEichelmann	http://execgraph.org/Hochtief_AG	0	1	2001	2014	13
http://execgraph.org/ThomasEichelmann	http://execgraph.org/V-Bank_AG	0	1	2007	2018	11
http://execgraph.org/ThomasEichelmann	http://execgraph.org/Wuestenrot_&_Wuerttembergische_AG	0	1	2012	2017	5

Table 5: Experience in board activities of members of Wirecards supervisory board

When analyzing the members of the supervisory board, it remains to be seen that there are two fundamental conspicuous characteristics. Firstly, the small size. As already mentioned, the supervisory board has always been smaller than at other DAX30 companies in the period from 2013 to 2019. It is difficult to say whether this is a decisive reason for the downfall of the company. However, it does seem striking.

On the other hand, the lack of experience of members of the supervisory board in other management and supervisory boards. The lack of experience of the supervisory board in board activities is unlikely to have facilitated the supervision of the board. Accordingly, it was certainly a right step to include Stefan Eichelmann with all his experience in the supervisory board. However, there is another factor. Due to the lack of experience in board activities, the members of the supervisory board also have comparatively few links to other members of the supervisory and management boards.

This could already be seen in the evaluations above. All these circumstances are indications that the supervisory board at Wirecard did not properly monitor the management board.

Table 6 gives an overview of the number of total connections, the number of board members and the connections per board member.

DAX Company	Connections	Board Members	Connections per Board Member
MTU Aero Engines	278	16	17,4
Wirecard AG	442	6	73,7
Vonovia SE	575	16	35,9
HeidelbergCement AG	700	19	36,8
Beiersdorf AG	777	19	40,9
Fresenius SE & Co. KG	941	20	47,1
Fresenius Medical Care KG	969	14	69,2
Linde Plc	1366	19	71,9
adidas AG	1494	21	71,1
Deutsche Börse AG	1575	24	65,6
SAP SE	1929	24	80,4
Münchener Rückversicherung	1952	30	65,1
Infineon Tech AG	2130	25	85,2
Merck KG	2188	21	104,2
Deutsche Post AG	2238	28	79,9
Volkswagen AG	2315	27	85,7
Deutsche Telekom AG	2325	27	86,1
Lufthansa AG	2343	31	75,6
E.ON SE	2501	25	100
Continental AG	2503	28	89,4
Deutsche Bank AG	2622	28	93,6
RWE AG	2689	24	112
Covestro AG	2760	16	172,5
BASF SE	3211	18	178,4
Allianz SE	3403	22	154,7
Henkel AG & Co. KG	3526	23	153,3
Bayer AG	3969	25	158,8
BMW AG	4578	27	169,6
Daimler AG	5071	29	174,9
Siemens AG	5701	27	211,1

Table 6: Overview of connections per company

The table gives a clear overview of the number of connections of the board members. It is particularly important to note that Wirecard does not have the least connections and that the value for connections per board member is not comparatively low. However, this is due to the fact that the connections are reduced to a few people. These include Stefan Eichelmann, for example.

5 Conclusion

In this paper we created a knowledge graph to analyze the people in the Wirecard scandal. These include mainly the members from the management and supervisory board. Basically, it can be summarized that the use of a knowledge graph works very well in this context. A similar evaluation with Excel, for example, would have hardly led to the same results. This is mainly due to the fact that the data in the knowledge graph could be queried on the one hand, but also visualized on the other. This led to evaluations that would otherwise not have been possible. This includes, for example, the analysis of industry connections as well as inter-board connections.

The results of these analyses lead to the following central theses.

- The members of the Board have had no experience in other board activities.
- The experience of the members of the supervisory board has varied greatly. There were members who already had experience. However, some members of the supervisory board also had no experience in board activities. In this context, the name of Thomas Eichelmann should be mentioned in particular, who has experience in board activities, but who only joined the supervisory board shortly before the company's failure.
- Very little was known about the members of Wirecard's management board. Accordingly, the data basis here was very small. This is also unusual. Especially since companies in Germany provide information about the members of the management board and the supervisory board on their websites. Here, the number and quality of the information can vary greatly, but the missing information, such as for Jan Marsalek, stood out among the members of the database.
- The members of Wirecard's management board have very few links with other members of the management and supervisory boards of other companies.

This is due on the one hand to the small database and on the other hand to the lack of experience in board activities.

In particular, the results of the analysis of the connections between the members of the management board and the supervisory board are difficult to evaluate. On the one hand, it is not desirable for networks such as the old boy network to exist. On the other hand, however, it can be assumed that a shared professional experience leads to a better assessment of another person. So, if you already know the qualities of a person, then you will probably also like to work with this person.

References

Bartz, T., Becker, S., Buschmann, R., Grozev, C., Hesse, M., Höfner, R., Hoppenstedt, M., Knobbe, M., Lehberger, R., Naber, N., Polonyi, M., Rosenbach, M., Schmid, F., Schulz, T., Seith, A., Traufetter, G., Wiedmann-Schmidt, W., Winterbach, C. (2020, July 18). *Auf der Jagd nach Dr. No*. Der Spiegel. https://blendle.com/i/der-spiegel/auf-der-jagd-nach-dr-no/bnl-derspiegel-20200718-1ef93f48bad?sharer=eyJ2ZXJzaW9uIjoiMSIsInVpZCI6InNhcmFoaGFja2wyMDAwIiwiaXRlbV9pZCI6ImJubC1kZXJzcGllZ2VsLTIwMjAwNzE4LTFlZjkzZjQ4YmFkIn0%3D.

Berners-Lee, T. (2009, June 18). *Linked Data - Design Issues*. W3. https://www.w3.org/DesignIssues/LinkedData.html.

Boyd, R. (2018). Wirecard AG: The Great Indian Shareholder Robbery. The Foundation for Financial Journalism. URL: https://ffj-online.org/2018/01/23/wirecard-ag-the-great-indian-shareholder-robbery/.

Davies, P. J. (2020, June 22). *Wirecard Says Missing Billions Don't Exist*. The Wall Street Journal. https://www.wsj.com/articles/wirecards-missing-2-billion-probably-doesnt-exist-board-says-11592802732.

Drescher, R. & Kirchner, C. (2008). *Wirecard beauftragt Ernst & Young.* Handelsblatt, 10–12.

DSW-Aufsichtsratsstudie 2018. (2018). DSW. https://www.dsw-in-fo.de/fileadmin/Redaktion/Dokumente/PDF/Presse/DSW_Pressekonferenz_Aufsichtsratsstudie_2018_-_Grafiken.pdf.

DSW-Aufsichtsratsstudie 2019. (2019). DSW. https://www.dsw-in-fo.de/fileadmin/Redaktion/Dokumente/PDF/Presse/DSW_Pressekonferenz_Aufsichtsratsstudie_2019_-_Grafiken.pdf.

DSW-Aufsichtsratsstudie 2020. (2020). DSW. https://www.dsw-in-fo.de/fileadmin/Redaktion/Dokumente/PDF/Presse/DSW_Pressekonferenz_Aufsichtsratsstudie_2020_-_en.pdf.

Ehrlinger, L., & Wöß, W. (2016). *Towards a Definition of Knowledge Graphs*. SEMANTICS. http://ceur-ws.org/Vol-1695/paper4.pdf.

Giesen, C., Kampf, L., Munzinger, H., Ott, K., & Willmroth, J. (2021, February 02). *Der Mann, der Wirecard jagte*. Süddeutsche Zeitung (SZ). https://www.sueddeutsche.de/wirtschaft/wirecard-dan-mccrum-1.5193416?reduced=true.

Grüll, P., Meyer-Fünffinger, A., Streule, J. & Wolf, S. (2020, December 07). *Die Suche nach den 1,9 Milliarden Euro*. Tagesschau. https://www.tagesschau.de/investigativ/br-recherche/wirecard-milliarden-versteckspiel-101.html.

Hübner, A. (2021, May 21). *Insolvenzverwalter - Wirecard-Milliarden haben nie existiert*. Reuters. https://www.reuters.com/article/deutschland-wirecard-id-DEKCN2D21SH.

Kejriwal, M. (2019). Domain-Specific Knowledge Graph Construction - What Is a Knowledge Graph?, Springer International Publishing. pp. 1–7.

KPMG. (2020). *Bericht über die unabhängige Sonderuntersuchung*. https://www.wirecard.com/uploads/Bericht_Sonderpruefung_KPMG.pdf.

Krahnen, J. P. & K. Langenbucher, K. (2020). *The Wirecard lessons: A reform proposal for the supervision of securities markets in Europe*. SAFE Policy Letter, Research Report. (88). https://www.econstor.eu/handle/10419/222230.

Lenz, H. (2020). Die Verantwortung des Abschlussprüfers zur Aufdeckung von Bilanzdelikten (Täuschungen, Vermögensschädigungen): Das Fallbeispiel Wirecard AG. KoR : internationale und kapitalmarktorientierte Rechnungslegung ; IFRS, 20(12).

Malcher, I., Schieritz, M. & Willeke, S. (2020). *Was hat dieser Mann mit Wirecard zu tun?* Die Zeit, 15–23. https://www.zeit.de/2020/41/wirecard-skandal-jan-marsalek-geldwaesche-karibik-accompong?utm_referrer=https%3A%2F%2Fwww.google.com%2F.

McCrum, D. (2015, April 27). *The House of Wirecard*. Financial Times. https://www.ft.com/content/534e7c4d-3101-3f6a-abc8-dc70beab35b7.

McCrum, D. (2019, October 15). *Wirecard's suspect accounting practices revealed*. Financial Times (FT). https://www.ft.com/content/19c6be2a-ee67-11e9-bfa4-b25f11f42901.

McCrum, D. & Palma, S. (2019, February 7). *Wirecard: inside an accounting scandal*. Financial Times. https://www.ft.com/content/d51a012e-1d6f-11e9-b126-46fc3ad87c65.

Paulheim, H. (2016). *Knowledge Graph Refinement - A Survey of Approaches and Evaluation Methods*. Semantic Web. http://semantic-web-journal.net/system/files/swj1167.pdf.

Peemöller, V. H., Krehl, H., Hofmann, S. & Lack, J. (2020). *Bilanzskandale* (3rd ed.). Erich Schmidt Verlag.

Peitsmeier, H. (2018, September 06). *Der Aufstieg des geheimnisvollen Zahlungsabwicklers*. Frankfurter Allgemeine Zeitung. https://www.faz.net/aktuell/wirtschaft/unternehmen/der-aufstieg-des-wirecard-chefs-markus-braun-15771197.html.

Rasch, M. (2020, June 23). *Aufstieg und Fall des Markus Braun – über einen Mann, der das Rampenlicht angeblich scheute*. Neue Züricher Zeitung. https://www.nzz.ch/wirtschaft/aufstieg-und-fall-des-markus-braun-ueber-einen-mann-der-das-rampenlicht-scheute-ld.1562809.

Robinson, I., Webber, J., & Eifrem E. (2013). *Graph databases*. O'Reilly Media, Inc.

Singhal, A. (2012, May 16). *Introducing the Knowledge Graph: things, not strings*. Google. https://blog.google/products/search/introducing-knowledge-graph-things-not/.

Slodczyk, K. (2020). *Wirecard: Ex-Aufsichtsrätin Tina Kleingarn bringt Markus Braun in Bedrängnis*. manager magazin. https://www.manager-magazin.de/unternehmen/wirecard-ex-aufsichtsraetin-tina-kleingarn-bringt-markus-braun-in-bedraengnis-a-c761bbbe-a261-4bf7-93f4-ea043a9e3193, zuletzt abgerufen am: 20.06.2021.

Storbeck, O., Palma, S. & McCrum, D. (2020, August 07). *Prosecutors suspect Wirecard was looted before collapse*. Financial Times. https://www.ft.com/content/c8acf321-7bc7-4348-99f6-b17e01085238.

Traufetter, G., Hesse, M., Böcking, D. & Bartz, T. (2020). *Wirecard-Skandal: Wirtschaftsprüfer EY und Aufsichtsbehörde machen sich gegenseitig Vorwürfe*. Der Spiegel. https://www.spiegel.de/wirtschaft/unternehmen/wirecard-skandal-wirtschaftspruefer-ey-und-aufsichtsbehoerde-machen-sich-gegenseitig-vorwuerfe-a-f011387f-8e19-4c82-99f7-037b27076b3f.

Velte, P. & Graewe, D. (2021). Reform der Corporate Governance nach dem Wirecard-Skandal. NWB Verlag.

Véron, N. (2020, June 30). *The Wirecard debacle calls for a rethink of EU, not just German, financial reporting supervision*. Bruegel. https://www.bruegel.org/2020/06/the-wirecard-debacle-calls-for-a-rethink-of-eu-not-just-german-financial-reporting-supervision/.

Appendix

Figure 4: Graph model

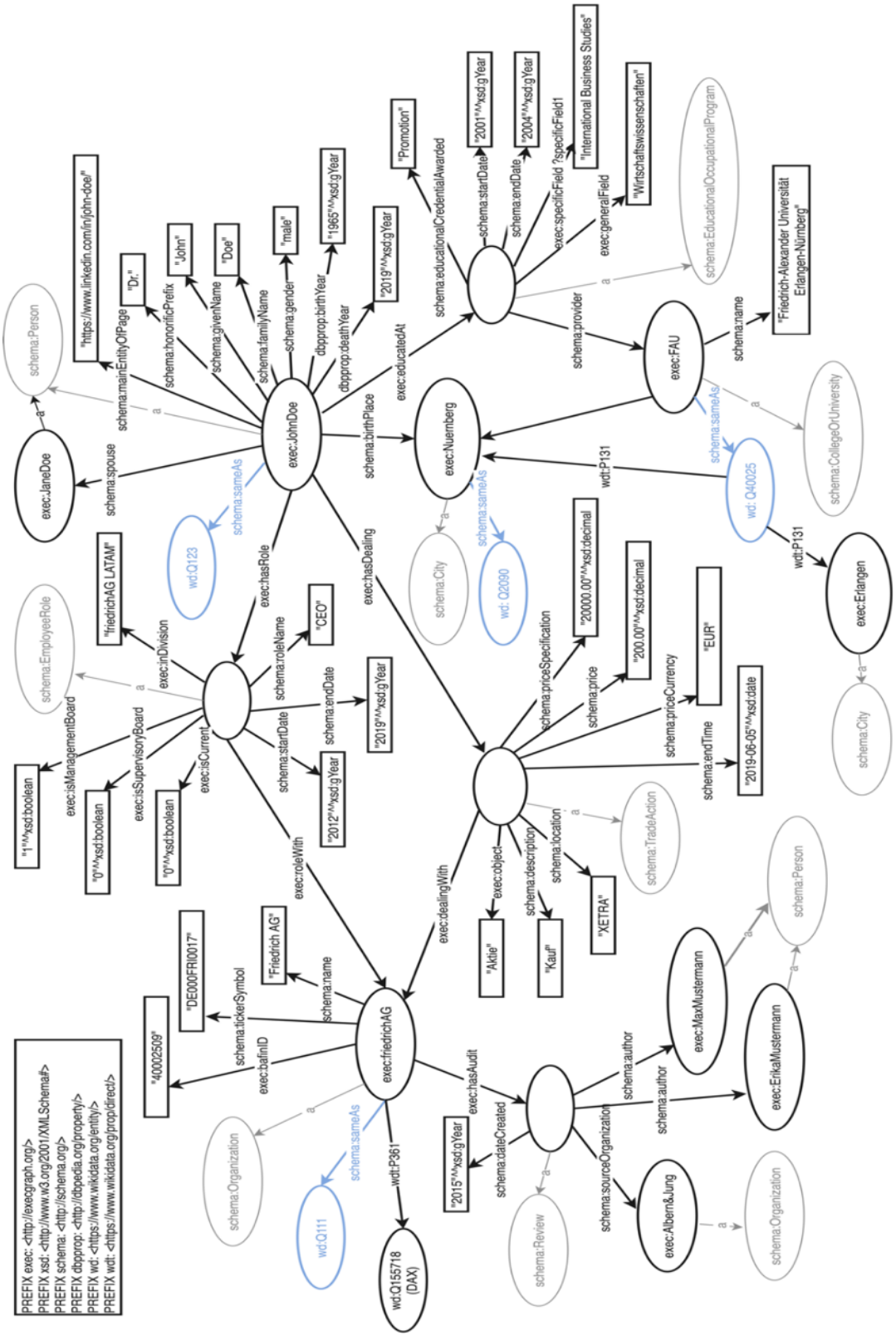


Table 4: Amount of the members of the supervisory board, management board and auditors of the DAX30 companies

Company	Supervisory board	Management board	Auditors
Adidas	16	6	2
Allianz	12	10	2
BASF	12	7	2
BMW	20	8	2
Bayer	20	7	2
Beiersdorf	12	8	2
Continental	20	8	2
Covestro	12	4	2
Daimler	20	8	2
Deutsche Bank	19	7	2
Deutsche Börse	16	6	2
Deutsche Post	20	8	2
Telekom	20	9	2
EON	20	5	2
Fresenius Medical Care	6	7	2
Fresenius	15	7	2
HeidelbergCement	12	7	2
Henkel	16	5	2
Infineon	16	4	2
Linde[1]	11	8	-
Lufthansa	21	6	2
Merck	16	5	2
MTU	12	4	2
Munich Re	20	9	2
RWE	20	2	2
SAP	18	8	2
Siemens	20	8	2
VW	20	8	2
Vonovia	12	4	2
Wirecard[2]	6	4	-
Sum	**480**	**197**	**56**

[1] It is not possible to find detailed information of the executives and auditors of Linde plc.
[2] We do not have any information about the auditors of Wirecard because the annual report for the year 2019 was not published.

Section B.6

WireGraph - Entwicklung eines Lernspiels mit Prosumentenumgebung zur Förderung der fachlichen und digitalen Kompetenz von Wirtschaftsstudierenden

(with Christian Fleiner, Goran Zvekan & Andreas Harth)

This project was partially (10.000 Euro) funded by the Innovation Fund Education of the Friedrich-Alexander-Universität Erlangen-Nürnberg.

Published in:

Hochschule 2031 - digital und nachhaltig?!
In: Gesellschaft für Informatik, Bonn (ed.): INFORMATIK 2021 - Computer Science & Sustainability 2021

Presented at:

51. Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2021 - Computer Science & Sustainability, Berlin, Germany

Accepted for presentation at:

Knowledge Graph Conference 2022, New York City, USA

A version of this paper has accepted for presentation at:

27th annual conference on Innovation and Technology in Computer Science Education (ITiCSE)

Contents – Section B.6

WireGraph - Entwicklung eines Lernspiels mit Prosumentenumgebung zur Förderung der fachlichen und digitalen Kompetenz von Wirtschaftsstudierenden

Abstract

Die Europäische Kommission warnt davor, dass die Entwicklung der Digitalisierung unzureichend in der Hochschullehre beachtet wird. Um die Vermittlung von digitaler Kompetenz in die Lehre miteinzubinden, eignet sich das von der Europäischen Kommission vorgeschlagene Referenzmodell der digitalen Kompetenz. Es besteht ein großer Forschungsbedarf, wie das Referenzmodell explizit in die Lehre übertragen werden kann. In diesem Beitrag wird das Konzept des Lernspiels WireGraph vorgestellt, mit dem Studierende der Wirtschaftswissenschaften die Arbeit mit Wissensgraphen erlernen, um den Wirecard-Skandal aufzuarbeiten. Dazu wird eine Prosumentenumgebung gestellt, mit der der Kompetenzbereich Erstellung digitaler Inhalte des Referenzmodells vollständig und über alle Leistungsniveaus hinweg vermittelt wird. Dabei erlernen Studierende nicht nur die Grundlagen von RDF und SPARQL, sondern können diese auch gezielt für ihre Ideen und weiteres Fachwissen einsetzen. Das Lernspiel WireGraph bildet mit der Konzeption die erste Stufe eines mehrstufigen Projektes. Die folgenden Schritte sind die Ausarbeitung des Lernspiels und anschließend die Anwendung und die Messung des didaktischen Erfolgs.

Keywords

Audit; Big Data; DigComp; Digitalisierung; Educational Game; Interaktive Lehre; Linked Data; Prosument; Wissensgraph; Wirecard

1 Motivation

Bereits 2017 hat die Europäische Kommission auf das "Missverhältnis zwischen den Kompetenzen, die in Europa benötigt werden, und den Kompetenzen, die vorhanden sind" als einer der vier großen Herausforderungen für die Europäische Union hingewiesen. Mit dem Vermerk, dass viele Absolventen:innen eine mangelhafte digitale Kompetenz aufwiesen, die aber dringend erforderlich sei, soll insbesondere in der Hochschulbildung mit geeigneten Maßnahmen und Mechanismen reagiert werden [Eu17]. In der Arbeitswelt wird bereits seit einiger Zeit deutlich, dass eine gute digitale Kompetenz nicht nur in den informationstechnischen Berufen vorausgesetzt wird[1]. Auch der Finanzbereich ist davon betroffen, wie beispielsweise das Berufsbild des Wirtschaftsprüfers [HS17]. Künstliche Intelligenz, Blockchain und Big Data werden als die Technologien genannt, die einen großen Einfluss auf Berufe des Rechnungswesens haben werden. [GHV19]. Das ist nachzuvollziehen, da erst durch den Einsatz dieser Technologien neue Möglichkeiten verfügbar wurden. So werden im Big Data-Kontext Wissensgraphen für Thematic Investing oder Risikoanalysen eingesetzt [Ni20]. Zudem konnten mit Hilfe von Wissensgraphen Verbindungen zwischen unautorisierten Informationen in der Panama-Papers-Affäre herstellt werden[2]. Aber auch andere Fachrichtungen, wie zum Beispiel Humanmedizin sind davon betroffen [KFT18]. Daher muss die Forderung an die Hochschulbildung, digitale Kompetenzen gezielt zu fördern, um nachhaltige Lehre zu erreichen, beachtet werden. Im Rechnungswesen identifizierten Cockcroft und Russell sechs forschungsrelevante Themengebiete von Big Data: Risiko, Sicherheit, Datenvisualisierung, Predictive-Analytics, Datenmanagement und Datenqualität [CR18]. Unter Datenmanagement ist zum Beispiel die Integration mehrerer Datenbänke zu verstehen und die Vermeidung von Redundanz. Datenqualität befasst sich mit dem Wahrheitsgehalt und der Inferenzmöglichkeit. Wissensgraphen sind eine Big Data-Technologie, die Mechanismen für Datenmanagement und -qualität liefert.

[1] https://www.econstor.eu/bitstream/10419/146097/1/843867167.pdf, letzter Zugriff am 14.04.2021.

[2] https://medium.com/analytics-vidhya/panama-papers-meet-knowledge-graph-c70652d98f92, letzter Zugriff am 14.04.2021.

Um auf diese Entwicklung zu reagieren, wird das Lernspiel *WireGraph* vorgestellt, mit dem Studierende der Wirtschaftswissenschaften die Arbeit mit Wissensgraphen erlernen, aber auch Fachwissen im Bereich Prüfungstechnik vermittelt bekommen.

Die Handlungsgrundlage für das Lernspiel bildet der Wirecard-Fall. Dieser ist insbesondere durch seine Aktualität interessant und bietet zudem teilweise noch ungeklärte Handlungsstränge. Als Maßstab für die digitale Kompetenz wird das Europäische Referenzmodell der digitalen Kompetenz für Bürger (DigComp) herangezogen [CVP17]. Das Lernspiel wird in einer Prosumentenumgebung bereitgestellt, in der Studierende die Möglichkeit haben, die unterschiedlichen Leistungsniveaus des Referenzmodells zu erreichen, indem sie Fachwissen nicht nur verstehen und anwenden, sondern auch weiterentwickeln. Der explizite Transfer von DigComp auf das Lehrdesign wurde als Forschungslücke erkannt [Si18] und wird mit der Einführung der Prosumentenumgebung und deren Ausgestaltung am Beispiel des Lernspiels adressiert.

In diesem Beitrag wird zunächst die digitale Kompetenz, ihr Referenzmodell und die Prosumentenrolle der Studierendenschaft generell beschrieben. Anschließend werden aktuelle Lehr-Lern-Konzepte beschrieben, die in den Lehrveranstaltungen eingesetzt werden, in denen *WireGraph* als unterstützendes Medium zukünftig verwendet werden soll. Die Prosumentenrolle wird dabei diesen gegenübergestellt. Abschließend wird das Lernspiel *WireGraph* aufgeführt und die Ausgestaltung der Prosumentenumgebung ausformuliert.

2 Einordnung der digitalen Kompetenz Motivation

2.1 Digitale Kompetenz - eine Definition

Es existiert keine einheitliche Definition der digitalen Kompetenz. In diesem Beitrag wird daher die Definition aus der *Empfehlung des Europäischen Rates zu Schlüsselkompetenzen für lebensbegleitendes Lernen* (2018) fortan verwendet [Ra18], die sich aus der Definition der *Computerkompetenz* [Eu06] herausbildete:

Digitale Kompetenz umfasst die sichere, kritische und verantwortungsvolle Nutzung von und Auseinandersetzung mit digitalen Technologien für die allgemeine und berufliche Bildung, die Arbeit und die Teilhabe an der Gesellschaft.

Sie erstreckt sich auf Informations- und Datenkompetenz, Kommunikation und Zusammenarbeit, Medienkompetenz, die Erstellung digitaler Inhalte (einschließlich Programmieren), Sicherheit (einschließlich digitales Wohlergehen und Kompetenzen in Verbindung mit Cybersicherheit), Urheberrechtsfragen, Problemlösung und kritisches Denken. […]

2.2 DigComp 2.1 - Referenzmodell der digitalen Kompetenz für Bürger

Das in Version 2.1 veröffentlichte Referenzmodell [CVP17] umfasst fünf Kompetenzbereiche, in denen Bürger Fähigkeiten und Kenntnisse aufweisen müssen, um in der heutigen und zukünftigen digitalen Gesellschaft selbstbestimmt agieren zu können. Die Kompetenzbereiche sollen im Folgenden kurz skizziert werden:

- **Informations- und Datenkompetenz**. Dieser Kompetenzbereich beinhaltet das Finden, Filtern, Interpretieren und das Verwalten von digitalen Inhalten, wie zum Beispiel die korrekte Verwendung einer Suchmaschine.
- **Kommunikation und Zusammenarbeit**. Hierzu zählt die bewusste Verwendung von Technologien zur Ermöglichung oder Verbesserung der zwischenmenschlichen Zusammenarbeit, wie zum Beispiel beim Einsatz von Videokonferenzanwendungen.
- **Erstellung digitaler Inhalte**. Hierzu gehört die Fähigkeit neue, digitale Inhalte zu erstellen und das Verständnis gesetzlicher Regelungen, wie zum Beispiel das (Kunst-)Urhebergesetz, im digitalen Umfeld.
- **Sicherheit**. In diesen Kompetenzbereich fällt das bewusste Schützen von Geräten, persönlichen Daten und allgemein der Privatsphäre im digitalen Umfeld.
- **Problemlösung**. Dazu gehört der Umgang mit technischen Problemen und der kreative Umgang mit Technologien, sowie die Erkennung von digitalen Kompetenzlücken zur nachhaltigen Weiterentwicklung des Referenzmodells.

Bemerkung Die in 2.1 genannte *Medienkompetenz* und das *kritische Denken* sind implizit im Referenzmodell enthalten.

In Bezug auf das Referenzmodell kann eine digitale Gesellschaft nur als nachhaltig gelten, solange deren Bürger in dieser selbstbestimmt handeln können.

Daraus ergibt sich der Auftrag an eine nachhaltige Hochschullehre, zu gewährleisten, dass Studierende nach Abschluss des Studiums die erforderlichen digitalen Kompetenzen dafür besitzen. Im Positionspapier des Kompetenzzentrums *Nachhaltige Universität* der Universität Hamburg[3] wird als Maßnahme genannt, dass Studierende die Möglichkeit erhalten und nutzen sollen, ihre Lernprozesse selbst zu gestalten. In *WireGraph* erlernen Studierende Wissensgraphen zu erstellen und abzufragen. Diese Kompetenz gehört zur Unterkategorie Programmierung und ist damit dem Kompetenzbereich *Erstellung digitaler Inhalte* zuzuordnen. Damit die Studierenden ihre Lerninhalte und -methoden allerdings selbst (mit-)gestalten können, ist es notwendig, dass das Wissen nicht nur gelernt, sondern auch angewendet wird. Wie es bereits im Referenzmodell aufgeführt wird, müssen eigene Inhalte erstellt und Ideen umgesetzt werden - aus dem konsumierenden Studierenden wird ein Prosument im digitalen Zeitalter.

2.3 Der Prosument im digitalen Zeitalter

Prosument ist ein Mischwort, bestehend aus den Begriffen *Produzent* und *Konsument*. Der Begriff *Prosumerismus* wurde erstmals im Jahre 1980 eingeführt [To80]. Im digitalen Umfeld ist derjenige Prosument, der sowohl digitale Inhalte selbst erstellt als auch Inhalte des gleichen Kontexts abruft bzw. verwertet. Die Erstellung und Verwertung kann dabei synchron oder asynchron stattfinden.

Die Prosumentenrolle wird im Referenzmodell implizit in zwei Dimensionen adressiert. (A) Auf der Makro-Ebene beinhaltet der Kompetenzbereich *Informations- und Datenkompetenz* hauptsächlich Anforderungen, die das korrekte Konsumieren von digitalen Inhalten betrifft. Die Produzentenrolle spiegelt sich im Kompetenzbereich *Erstellung digitaler Inhalte* wider. (B) Zusätzlich erfolgt eine zweite Trennung zwischen Konsument und Produzent innerhalb jedes Kompetenzbereichs auf der Mikro-Ebene in Folge der acht Leistungsniveaus [CVP17]. So ist es erforderlich, dass jeder zunächst erprobte Methoden und Inhalte der Kompetenzbereiche versteht und anwenden kann (Konsument; Stufe 1 – 4), bevor neue Inhalte gelehrt und hinzufügt werden (Produzent; Stufe 5 – 8).

[3] https://www.nachhaltige.uni-hamburg.de/downloads/2015-04nachhaltigkeitskonzept-knu-team2-endfassung.pdf, letzter Zugriff am 14.04.2021.

Eine Gegenüberstellung der Leistungsniveaus und der Prosumentenrolle wird in Tabelle 1 dargestellt. Die Prosumentenrolle ist dabei in jedem Leistungsniveau vertreten, doch wird sie für die Nachhaltigkeit umso relevanter, desto mehr Produzentenanteile sie einnimmt. Besonders im Hinblick auf die forschungsorientierte Lehre sollte das Erreichen des achten Leistungsniveaus angestrebt werden.

Leistungsniveau	**Beispiel**	**PR**
1 – Einfache Aufgaben mit Hilfe ausführen	MP mit Supervisor schreiben/versenden	K
2 – Ein. Aufgaben teilw. mit Hilfe ausführen	MP mit Rückfragen schreiben/versenden	
3 – Einfache Probleme selbstständig lösen	Formatierung-/Anhangsfunktion nutzen	
4 – Komplexe Probleme selbstständig lösen	File-Hosting-Dienst für Anhänge nutzen	
5 – Geeignete Lösungswege kommunizieren	Instant-Messenger einführen	P
6 – Besten Lösungsweg kommunizieren	Instant-Messenger für MP optimieren	
7 – Lösungswege verbessern	API des Instant-Messengers nutzen	
8 – Kompetenzbereich erweitern	Sprach-/Texterkennung integrieren	

Tabelle 1: Gegenüberstellung der Leistungsniveaus des Referenzmodells und der Prosumentenrolle im Kompetenzbereich *Erstellung digitaler Inhalte*.
Anm.: Als Beispiel dient ein Szenario, in dem ein Mitarbeiter ein Meetingprotokoll (MP) per E-Mail an sein Team senden soll. (PR - Prosumentenrolle; K - Konsumentenrolle; P - Produzentenrolle).

Um eine nachhaltige Hochschullehre zu gewährleisten, muss die Prosumentenrolle der Studierenden im Detail verstanden werden. Die in Abschnitt 1 erwähnte Forderung an die Hochschullehre bezieht sich auf die Vermittlung wichtiger digitaler Kompetenzen nach (A), die zwischen Studiengängen unterschiedlich gewichtet werden können. So werden Informatikstudierende im digitalen Umfeld öfter die Produzentenrolle einnehmen als Medizinstudierende. Viel bedeutender für eine nachhaltige Hochschullehre ist allerdings, dass Studierende die Möglichkeit erhalten, jede gelehrte digitale Kompetenz über alle Leistungsniveaus nach (B) kennenzulernen, sodass Studierende und Absolventen:innen bei der voranschreitenden Digitalisierung mitwirken können.

3 Aktueller Stand der eingesetzten Lehr-Lern-Konzepte

Eine Übersicht bestehender Lehr- und Lernformate präsentiert Kuhn et al. [KFT18]. Zwar bezieht sich die Übersicht auf die medizinische Ausbildung, kann aber generisch für die Hochschullehre betrachtet werden.

Das Lernspiel wird zukünftig als unterstützendes Medium in zwei Lehrveranstaltungen eingesetzt: Eine Lehrveranstaltung über Wissensgraphen und einer Lehrveranstaltung der Prüfungstechnik. Hier werden abgesehen von Übungen, digitale Präsentations- und Videodateien eingesetzt, die für Studierende auf einer E-Learning-Plattform abrufbar sind. Zusätzlich wird das Lernspiel *AuditSim* in der Prüfungstechniklehre eingesetzt. Für diese Lehr-Lernkonzepte wird nun die Möglichkeit einer Prosumentenumgebung analysiert:

- *Digitale Präsentations- und Videodateien* werden meist von den Dozierenden bereitgestellt, wodurch Studierende vornehmlich auf der Makro-Ebene die Konsumentenrolle einnehmen. In Praktika, Seminaren und Abschlussarbeiten sind Präsentationen und somit die Erstellung digitaler Präsentationsdateien für Studierende oftmals verpflichtend. Wenn diese digitale Kompetenzausprägung gelernt wird, sollte die Erreichung des achten Leistungsniveaus angestrebt werden.

  Das bedeutet nicht nur die Verwendung bestehender Vorlagen, sondern auch die Erarbeitung eigener Elemente, die zum Beispiel in Form von Designs oder kreativer Elementkombinationen erfolgen können. Nach diesem Ansatz ist es also für den Studierenden nicht vorteilhaft, wenn Universitäten Präsentationsvorlagen bereitstellen, da so die Möglichkeit der Produktion entfällt.

- *Lernspiele* fördern das kritische Denken und die Fähigkeit kreativ Probleme zu lösen [Jo14]. Sie können als didaktisches Mittel in der Hochschullehre als Teil eines Kurses eingesetzt werden [Jo11]. Dabei ist wichtig, dass Lernspiele in engen Bezug zu den vermittelnden Lerninhalten stehen [ZML12]. Abgesehen von *AuditSim*, wird auch das Lernspiel *SQL Island* als repräsentatives Programmierungs-Lernspiel betrachtet:

  - In *AuditSim*[4] schlüpfen die Teilnehmer in die Rolle eines Wirtschaftsprüfers. Ein virtueller Manager stellt Aufgaben und andere Ressourcen, in Form von

[4] https://www.pw.rw.fau.de/studium-lehre/master/schluesselqualifikation-fact/, letzter Zugriff am 14.04.2021.

Dokumenten oder Videos, zur Verfügung. Die Bearbeitung und Bewertung der Aufgaben finden in einem unabhängigen Textverarbeitungsprogramm statt. Dadurch ist ein Spielleiter erforderlich. Studierende sind hierbei Konsumenten der *Informations- und Datenkompetenz*. Allerdings entfällt die Informationssuche. Studierende haben keine Möglichkeit die Produzentenrolle einzunehmen. Dafür müsste es ihnen freistehen, Such- und Evaluationsmethoden selbst zu bestimmen. Neue Suchmöglichkeiten könnten durch die Anwendung von SQL oder SPARQL ermöglicht werden, durch die Datenbanken abgefragt werden können.

- *SQL Island*[5] ist ein browserbasiertes Lernspiel, um „SQL-Anfragen zu verstehen und zu formulieren" [SD15]. SQL ist als Programmiersprache dem Kompetenzbereich *Erstellung digitaler Inhalte* zuzuordnen. Die Lernaufgaben beziehen sich dabei auf die *Data Manipulation Language* (DML) von SQL, mit der relationale Datensätze geändert und abgefragt werden können. *SQL Island* wird nach Aussage der Autoren an Hochschulen „in Datenbankvorlesungen als unterstützende Übung eingesetzt" [SD15].

  Das Lernziel dieses Spiels ist es Grundlagen zu vermitteln. Studierende können daher ein Leistungsniveau bis zur 3.Stufe hinsichtlich SQL-Kenntnissen erreichen. Um Prosument zu werden, müsste *SQL Island* es Studierenden zusätzlich ermöglichen, die Datenbasis zu erweitern, eigene *Quests* zu definieren und die Ergebnisse mit Kommilitonen zu besprechen.

4 WireGraph - Wissensgraphen spielerisch lernen

4.1 Handlung

Der im Jahr 2020 aufgedeckte Wirecard-Skandal, der bis heute noch nicht vollständig aufgeklärt ist, ist thematisch dem Finanzbereich zuzuordnen, da das in der digitalen Zahlungsabwicklung spezialisierte Unternehmen eine Fehlsumme in Höhe von 1,9 Milliarden Euro mit nicht vorhandene Treuhandkonten in Asien verschleierte[6], was auch unter Studierenden bekannt sein sollte. In *WireGraph* übernimmt der Spieler die

[5] https://sql-island.informatik.uni-kl.de, letzter Zugriff am 14.04.2021.

[6] https://www.ft.com/content/284fb1ad-ddc0-45df-a075-0709b36868db, letzter Zugriff am 14.04.2021.

Rolle eines Investigativjournalisten, der die wichtigsten Meilensteine der Skandalaufdeckung durchläuft. Dies ist auch deshalb realitätsnah, da Dan McCrum, Journalist der Financial Times, einen wichtigen Beitrag zur Aufdeckung des Skandals beitrug[7]. Die Handlung orientiert sich dabei am Krimi-Genre, u.a. auch weil dieses das populärste Genre der Belletristik ist[8]. Der Spieler muss daher Personen von Interesse befragen oder Emails durchleuchten, um darauf aufbauend einen Graphen zu bauen oder den bestehenden Graphen mit gewonnenen Indizien als Filter abzufragen. Damit würde sich das Spiel, wie in einem Kriminalroman, im weiteren Verlauf des Geschehens weiterentwickeln. Zum einem erfährt der Spieler mehr über den Wirecard-Skandal und zum anderen wird der Spieler mit neuen Herausforderungen konfrontiert, indem die Komplexität der IT-Inhalte ansteigt.

4.2 Zielgruppe und Lernziele

Die Zielgruppe des Spiels sind Masterstudierende der Betriebswirtschaftslehre und der Wirtschaftsinformatik, da die Modulhandbücher beider Gruppen an der Autorenuniversität bereits Vertiefungen zum Thema *Big Data* beinhalten.

Zudem konnten beide Gruppen bereits Fachwissen ansammeln, das ihnen helfen könnte, kreative Produktionen an dem Lernspiel vorzunehmen (wie zum Beispiel zum Thema Vertrags- oder Urheberrecht). Abschließend ermöglicht die geltende Studienprüfungsordnung beiden Gruppen fortgeschrittene Kurse im Bereich Wissensgraphen zu besuchen, sodass es ihnen offensteht, sich weiter in das Thema zu vertiefen. Die Lernziele orientieren sich an dem Referenzmodell der digitalen Kompetenz und sind wie folgt definiert:

- Teilnehmer verstehen das grundlegende Arbeiten mit Wissensgraphen.
- Teilnehmer können SPARQL-Abfragen schreiben, um Informationen auf Wissensdatenbanken, wie Wikidata[9], abrufen zu können.
- Teilnehmer können selbstständig Graphen auf Basis von Turtle- oder RDF/XML-Ausdrücken erstellen.

[7] https://www.ft.com/content/745e34a1-0ca7-432c-b062-950c20e41f03, letzter Zugriff am 14.04.2021.

[8] https://www.splendid-research.com/de/studie-buecher.html, letzter Zugriff am 14.04.2021.

[9] https://www.wikidata.org/, letzter Zugriff am 14.04.2020.

- Teilnehmer verstehen den Wirecard-Skandal und ihre Zusammenhänge.
- Teilnehmer verstehen den Aufbau und die Möglichkeiten der Lernspiel-Vorlagen und können alternative Lernszenarien entwickeln.

4.3 Umsetzung der Prosumentenumgebung

Das Lernspiel wird in Python und dem Kivy-Framework programmiert und ist derzeit noch in der Entwicklung. Das Spiel ist in Szenen gegliedert, wobei jede Szene eine andere Spielmechanik aufweisen kann, wie zum Beispiel die einer Visual-Novel- oder einer Point&Click-Spielweise. Dies soll mit Abbildung 1 veranschaulicht werden.

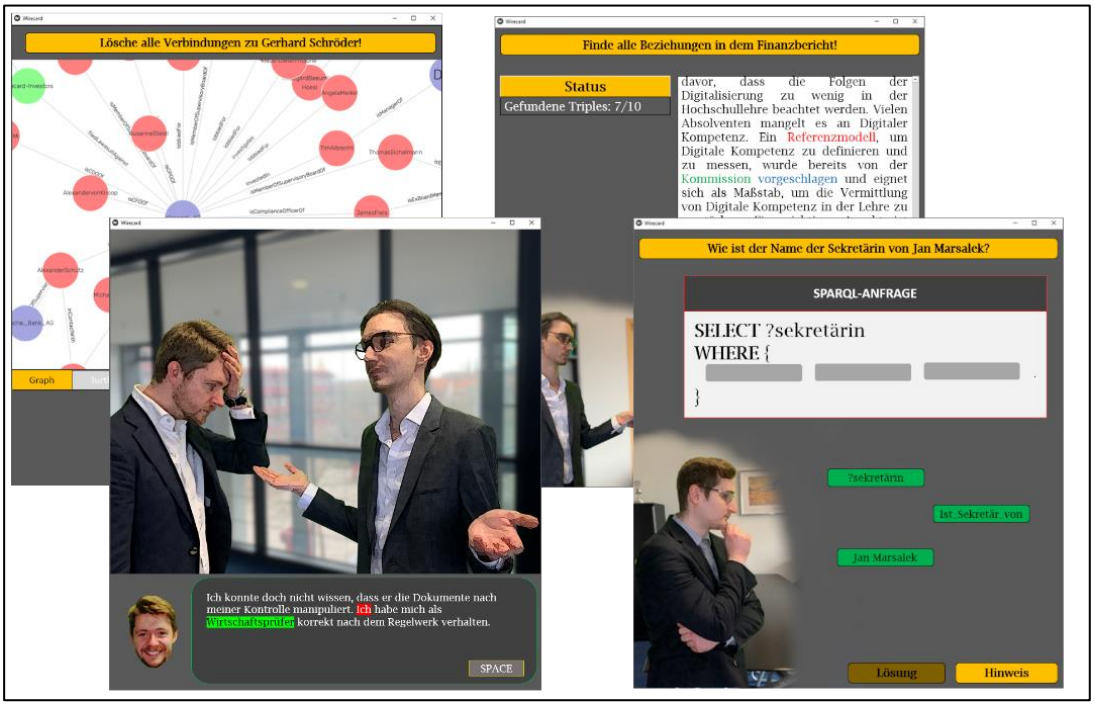


Abb. 1: WireGraph-Konzeptbilder.
Anm.: WireGraph-Konzeptbilder zeigen (v.l.n.r.): Visuelles Graphenmodellierungswerkzeug, Dialog- und Domentenanalyse, um Graphen zu erweitern, SPARQL-Abfrage (Drag&Drop).

Aktuell existieren Testszenen, mit denen die erforderlichen Interaktionselemente bestätigt wurden. Jede verfügbare Spielmechanik steht durch eine Szenen-Vorlage zur Verfügung, das mit Hilfe von RDF-Ausdrücken konfiguriert wird. Damit ist gewährleistet, dass die Teilnehmer ihr konsumiertes Wissen direkt für die Produktion eigener Inhalte verwenden können.

Python-Kenntnisse sind daher nicht erforderlich, können aber genutzt werden, um neue Vorlagen zu erstellen.

Das Hauptprojekt, sowie Projekte der Studierenden, werden auf GitHub gehostet. Das hat den Vorteil, dass Studierende ihre Projekte als Referenz potenziellen Arbeitgebern vorlegen und eine geeignete Softwarelizenz wählen können. Vergleichbar mit der Pygame-Projektseite[10] wird eine Übersicht aller Projekte gehostet, die auf das jeweilige GitHub-Projekt verweist, damit Studierende einen schnellen Überblick über die verfügbaren Lehrmaterialien bekommen. Die Prosumentenumgebung ergibt sich daher aus verfügbaren Spielen zum Erlernen der Thematik, der Übersichtsseite mit Kommentarfunktion zum Austausch und den Szenen-Vorlagen zur Erstellung neuer Lernspiele. Dadurch deckt die Prosumentenumgebung alle vier Unterpunkte des Kompetenzbereichs *Erstellung digitaler Inhalte* ab: Entwicklung von digitalem Inhalt, Überarbeitung von digitalem Inhalt, Lizenzvergabe und Programmieren.

4.4 Didaktische Einbettung und geplante Evaluation

Um die digitale Kompetenzausprägung hinsichtlich Wissensgraphen nachhaltig und vollständig zu vermitteln, muss das Lernspiel so ausgestaltet sein, dass Studierende die acht Leistungsniveaus des Referenzmodells in linearer Reihenfolge erreichen können [CVP17]. Das bedeutet, dass zunächst ein Verständnis aufgebaut, dann Wissen bewusst angewendet und abschließend kreativ erweitert werden soll. Für die Prosumentenrolle wird also zunächst die Konsumentenrolle und anschließend die Produzentenrolle eingenommen. Um die Prosumentenumgebung für *WireGraph* abzubilden, müssen drei Phasen durchlaufen werden:

1. Konsumption. Studierende spielen das Lernspiel.
2. Reflektion. Studierende diskutieren und evaluieren das gespielte Lernspiel mit anderen Kommilitonen.
3. Produktion. Auf Basis der Reflektionsphase erweitern bzw. verbessern Studierende das Spiel, um das Lernziel besser zu adressieren.

Die Entwicklung von *WireGraph* ist im Oktober 2021 abgeschlossen. Ab dem Wintersemester 2021/22 wird *WireGraph* in der Prüfungstechnik-Lehrveranstaltung als informatives Medium eingesetzt.

[10] https://www.pygame.org/tags/all, letzter Zugriff am 14.04.2021.

Ab dem Sommersemester 2022 wird *WireGraph* auch als unterstützendes Medium im Rahmen der Wissensgraph-Lehrveranstaltung für Studierende eingesetzt. Das Lernspiel kann von den Studierenden explizit dazu genutzt werden, um das Schreiben von RDF-Ausdrücken und SPARQL-Abfragen zu üben (Konsumption), da diese Fähigkeit für ein erfolgreiches Bestehen der Abschlussprüfung erforderlich ist. Am Ende beider Veranstaltungen wird den Studierenden ein Evaluationsbogen bezüglich *WireGraph* gegeben. Im Wintersemester 2022/23 wird Studierenden ein Praktikum angeboten, in dem sie *WireGraph* erweitern oder eigene Lernszenarien mit Hilfe der Vorlagen umsetzen (Produktion). Mit diesen Studierenden wird ein qualitatives Interview durchgeführt. Eine quantitative Analyse ist zunächst nicht vorgesehen, wird aber momentan nicht ausgeschlossen. Mit dem Wintersemester 2022/23 endet die erste Iteration.

Folgende Forschungsfragen (FF) werden dabei für *WireGraph* aufgestellt und sollen nach der ersten Iteration beantwortet werden:

FF1 Stellt die Prosumentenumgebung von *WireGraph* ein geeignetes Mittel dar, den Kompetenzbereich *Erstellung digitaler Inhalte* für Wissensgraphen vollständig abzubilden (Makro-Ebene)?

FF2 Erreichen Studierende durch die Prosumentenumgebung von *WireGraph* alle acht Leistungsniveaus des Kompetenzbereichs *Erstellung digitaler Inhalte* (Mikro-Ebene)?

FF3 Zeigen Studierende durch die Prosumentenumgebung eine höhere Bereitschaft sich mit dem Thema auseinanderzusetzen, weil sie beispielsweise digitale Inhalte (in Zusammenarbeit oder in Konkurrenz mit Kommilitonen) erstellen können (Peer-Motivation [Ha04]; Produktion)?

5 Zusammenfassung und Ausblick

Es besteht ein großer innerer und äußerer Druck für die Hochschullehre digitale Kompetenzen nachhaltig zu vermitteln. Unter anderem wird *Big Data* als zukunftsweisendes Thema für den Finanzbereich genannt, weshalb Studierende der Wirtschaftswissenschaften sich mit Technologien im Big Data-Umfeld auseinandersetzen müssen, um auf dem Arbeitsmarkt konkurrenzfähig zu bleiben. Wissensgraphen sind eine dieser Technologien und die Arbeit mit diesen soll mit dem vorgestellten Lernspiel *WireGraph* erlernt werden. Um die digitale Kompetenz einzuordnen, schlägt die Europäische Kommission ein geeignetes Referenzmodell der digitalen Kompetenz mit fünf Kompetenzbereichen und acht Leistungsniveaus vor, deren Anwendung auf Lehr-Lern-Konzepte noch Forschungslücken aufweist. Zusätzlich wird das Konzept der Prosumentenumgebung vorgeschlagen, die für das Lernspiel *WireGraph* erstmals umgesetzt wird, um die Kompetenz der Wissensgrapharbeit im Rahmen des Kompetenzbereichs *Erstellung digitaler Inhalte* vollständig und über alle Leistungsniveaus hinweg abzudecken. Dabei müssen Studierende die Prosumentenrolle einnehmen, in der sie digitales Fachwissen nicht nur konsumieren, sondern auch weiterentwickeln. Studierende aufarbeiten in *WireGraph* mit Hilfe von RDF und SPARQL den Wirecard-Skandal. Anschließend haben sie die Möglichkeit mit Hilfe der Szenen-Vorlagen eigene Ideen und Fachwissen in einem eigenen Wissensgraph-Lernspiel umzusetzen. *WireGraph* wird erstmalig im Wintersemester 2021/22 eingeführt und nach einem Jahr wird *WireGraph* für alle drei Phasen *Konsumption*, *Reflektion* und *Produktion* evaluiert worden sein. Dazu wurden drei Forschungsfragen aufgestellt, denen nachgegangen wird, um erste Ergebnisse zu erhalten und damit die Bedeutung der Prosumentenumgebung für die zukünftige Hochschullehre zu ermitteln.

Literaturverzeichnis

[CR18] Cockcroft, S.; Russell, M.: Big Data Opportunities for Accounting and Finance Practice and Research. Australian Accounting Review, 28(3):323–333, 2018.

[CVP17] Carretero, S.; Vuorikari, R.; Punie, Y.: The digital competence framework for citizens. Publications Office of the European Union, 2017.

[Eu06] Europäisches Parlament und Rat: Empfehlung vom 18. Dezember 2006 zu Schlüsselkompetenzen für lebensbegleitendes Lernen. Amtsblatt der Europäischen Union, L394:10 – 18, Dezember 2006.

[Eu17] Europäische Kommission: Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen- über eine europäische Erneuerungsagenda für die Hochschulbildung. COM, 247, Mai 2017.

[GHV19] Gulin, D.; Hladika, M.; Valenta, I.: Digitalization and the Challenges for the Accounting Profession. Proceedings of the ENTRENOVA - ENTerprise REsearchInNOVAtion Conference, 5(1):428–437, Oct. 2019.

[Ha04] Hancock, D.: Cooperative Learning and Peer Orientation Effects on Motivation and Achievement. The Journal of Educational Research, 97(3):159–168, 2004.

[HS17] Henselmann, K.; Scherr, E.: Auswirkungen der Digitalisierung auf den Berufsnachwuchs in der Wirtschaftsprüfung. In (Baldauf, J.; Graschitz, S., Hrsg.): Theorie und Praxis aus Rechnungswesen und Wirtschaftsprüfung, S. 231–241. LexisNexis Verlag, Wien, 2017.

[Jo11] Johnson, L. et.al.: The 2011 Horizon Report. Bericht, EDUCAUSE, 2011.

[Jo14] Johnson, L. et.al.: Horizon report Europe: 2014 schools edition. Bericht, EDUCAUSE, 2014.

[KFT18] Kuhn, S.; Frankenhauser, S.; Tolks, D.: Digitale Lehr-und Lernangebote in der medizinischen Ausbildung. Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz, 61(2):201–209, 2018.

[Ni20] Nigam, V. et.al.: A Review Paper On The Application Of Knowledge Graph On Various Service Providing Platforms. In: 2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence). S. 716–720,2020.

[Ra18] Rat der Europäischen Union: Empfehlung vom 22. Mai 2018 zu Schlüsselkompetenzen für lebenslanges Lernen. Amtsblatt der Europäischen Union, C189:1–13, Mai 2018.

[SD15] Schildgen, J.; Deßloch, S.: SQL-Grundlagen spielend lernen mit dem Text-Adventure SQL Island. In (Seidl, T.; Ritter, N.; Schöning, H.; Sattler, K.; Härder, T.; Friedrich, S.; Wingerath, W., Hrsg.): Datenbanksysteme für Business, Technologie und Web (BTW 2015). Gesellschaft für Informatik e.V., Bonn, S. 687–690, 2015.

[Si18] Sicilia, M. et.al.: Digital Skills Training in Higher Education: Insights about the Perceptions of Different Stakeholders. In: Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality. TEEM'18, Association for Computing Machinery, New York, NY, USA, S. 781–787, 2018.

[To80] Toffler, A.: The third wave, Jgg. 484. Bantam books New York, 1980.

[ZML12] Zender, R.; Moebert, T.; Lucke, U.: RouteMe - Routing in Ad-hoc-Netzen als pervasives Lernspiel. In (Desel, J.; Haake, J.; Spannagel, C., Hrsg.): DeLFI 2012: Die 10. E-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V. Gesellschaft für Informatik e.V., Bonn, S. 201–212, 2012.

Chapter C

Conclusions

Contents – Chapter C

1 Summary of main findings

1.1 Zielführende Betriebsprüfungen durch Nutzung von Alternative Data?

The first paper analyzes how financial authorities can use alternative data to detect discrepancies in tax declarations or source of risks.

With regards to the tax proceedings, the financial authorities have to select companies for the audit and identify areas prone to tax risks. Traditionally, financial authorities use tax assessment data, comparison of key figures or data from the so-calles “E-Bilanz”. Other points of reference can also be the reporting obligations of the so-calles “Country-by-Country-Reporting” and the reporting obligations of tax plannings. In most cases, the information of taxpayers is unaudited. Hence, the use of alternative data is reasonable.

The main findings of this work can be summarized in the following theses:

- There are already extensive technical possibilities for the use of alternative data that are far from being exhausted. Apparent possibilities for use include consistency checks (of data currently available internally on taxpayers with external information from alternative data) on the one hand, and the identification of risk areas where there are "warning signals" (red flags) for inadvertently or even intentionally incorrect or missing tax returns on the other. With regards to consistency checks, no real-world evaluations can be shown as the data is subject to tax secrecy. However, three exemplary analyses illustrate the potential for obtaining “warning signals” from alternative data (Hofmann, 2008).
- However, the examples also illustrate that the data are often "fuzzy". Problems for the evaluation arise due to incomplete data, different spellings, typing errors, identical names of different persons (many data records do not have a unique identifier/primary key) and many others.
- The linking of several different data sets is also hindered by the fact that – even if they each have a unique identifier – these identifiers/primary keys are not the same. Mutual assignment is therefore only possible more or less precisely via auxiliary variables (such as the combination of name and date of birth).
- The upgrading of the transparency register to a full register is an important step. However, transitional periods still apply at present.

The fundamental problem remains that the entries are based on self-disclosures by the entitled parties. At least transparency has been increased.

- It is regrettable that already within Germany the public registers do not fulfill the conditions for "Open Data" in the sense of the Open Knowledge Foundation. As a rule, only individual data sets can be queried, which results in considerable costs (Monk et al., 2019).
- Corresponding public register data would be required at least throughout the EU. This would make data collection much easier and more targeted for the tax administrations of all countries.
- The less open data there is, the more time-consuming and potentially error-prone the collection and processing of the data is. However, for certain data, the purchase of already prepared data collections from commercial providers would be an alternative. This industry is growing rapidly. If it is not possible to buy complete data sets, Robotic Process Automation (RPA) provides a way for mass automated individual queries (Moffitt et al., 2018).
- For the collection, integration and analysis of alternative data, it is a good idea to use established and proven international standards. In the authors' opinion, the most suitable is a flexible graph database in the version of an RDF store.
- Three clear trends can be identified in a comparison over time: First, the amount of potentially available alternative data is growing permanently and exponentially. Second, there are increased efforts to make government data available as open data. And third, the idea of "linked" Linked Open Data is slowly taking hold. However, this potential must also be leveraged. This requires the will to set up an appropriate system of people and technology.

1.2 Analyzing the quality of iXBRL company accounts in the UK

This paper closes an important research gap by investigating the quality of SME iXBRL filings in the UK. The quality of the iXBRL filings is especially important as they represent an essential resource for companies. In addition, false information can mislead investors in their decision-making process.

The main findings can be summarized as follows:

- Interestingly, 81,20 % of the companies are micro companies. Most of the companies (60,09 %) have less than 100.000 GBP, but more than 10.000 GBP in terms of their total assets. Most of the companies are in the south of the UK. Almost every second iXBRL filing is created using IRIS Accounts Production Software (built-in) (iXBRL Tagging Features, 2019).
- My results show that the most common error which occurs in 9.639.583 tags is a missing field value. For example, it is likely that a company has forgotten to tag a value or has not tagged it properly. The second most common error which was found in 3.044.048 tags is an incorrect statement of time. The third most common error which appears in 1.038.675 tags is the wrong structure of the tags. This error can occur in various forms, for instance the structure of the tags may have been interchanged.
- However, the results show that the quality of the iXBRL accounts is very good and that most iXBRL filings do not include errors. This result is especially interesting. The UK is the only country where the filing of iXBRL accounts of SME companies is mandatory. In addition, this study focuses on SME companies and proves that the company size does not necessarily determine the quality of the iXBRL filing. Therefore, the data could be used as a basis for many investigations.

1.3 A Knowledge Graph from UK Financial Statements

The third paper analyzes UK financial statements with the help of knowledge graphs. The graph also connects to the existing knowledge bases, namely DBpedia, Open Corporates, and Wikidata. This way, the graph can grow over time and continuously generate knowledge.

The main findings can be summarized as follows:

- In total, there are seven steps to create the knowledge graph. To build our graph and to develop the ontologies, we reuse various existing ontologies describing the company and officer information. This step-by-step instruction can be used to create other knowledge graphs with some modifications based on datasets and data sources (OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax, 2012).
- In 2020, micro companies account for 99.3 percent of the business population in the UK (Business population estimates for the UK and the regions 2020, 2020). According to the result, the average profit is around GBP 63.000.
- The companies of the dataset are widely distributed across the UK. Most of the companies are in the south of the UK and London. Statistics show that 35 percent of the businesses are in those areas, specifically 1,1 million businesses in London and 0,93 million businesses in the Southeast of England (Business population estimates for the UK and the regions 2020, 2020).
- The information of the average number of employees is divided into five groups (starting from zero employees to more than 250 employees), like the statistical data provided by the UK government (Rhodes, 2018).

1.4 How to visualize relationships between supervisory board and manage ment board members and auditors using Neo4j

This paper illustrates how relationships between supervisory board members, management board members, and auditors can be visualized using the graph database Neo4j. The sample covers DAX30 companies in 2019 and includes 480 members of the supervisory board, 197 members of the management board and 56 auditors. The topic is of interest for two reasons. First, the special institutional setting follows a so called "principle of separation". The supervision and the management of corporations should be clearly separated (Section 111 (1) GSCA). In the past and the present, there are various examples of multiple mandates and personal ties between the members. Second, those relationships are not only difficult to identify, but also to analyze. Therefore, we have chosen a graph database (Neo4j) to detect, visualize and examine personal ties.

The main findings are summarized as follows:

- Knowledge graphs demonstrate their ability to analyze connections between persons of interest in this context.
- Neo4j enables the visualization of connections of members of the management board and the supervisory board.
- Neo4j enables even people not familiar with IT to analyze and visualize Knowledge graphs (Robinson et al., 2013).
- There are several examples for connections between members of the management board, members of the supervisory board and auditors. These connections can be between members of the same company, but also between members from different companies (Gröls, 2011; Oehmichen, 2011).

1.5 What can we learn from Knowledge Graphs? A Wirecard perspective

The fifth paper analyzes the connections of management and supervisory board of Wirecard by using a knowledge graph. The paper shows the advantages of using a knowledge graph instead of tools such as MS Excel. This led to evaluations that would otherwise not have been possible. This includes, for example, the analysis of industry connections as well as inter-board connections.

The main findings can be summarized as follows.

- The main advantage of using knowledge graphs to evaluate management and supervisory board connections lies in the fact that the data in the knowledge graph can be both queried and visualized at the same time (Singhal, 2012).
- The members of the management board have had no prior experience in other board activities.
- The experience of the members of the supervisory board varies as some members have no prior experience with supervisory board mandates, whereas other members have had prior board mandates.
- As there is very little information on the members of Wirecard's management board publicly available, the data basis is very small. In comparison with other DAX30 companies, this is rather unusual. All DAX30 companies provide comprehensive information on the members of the management and supervisory board on their websites (DSW 2018; DSW 2019; DSW 2020).
- The information on the members of the management board varies greatly. Especially with regards to Jan Marsalek, there was little information available on Wirecard's website (Grümmer, 2021).
- The members of Wirecard's management board only have a few links with other members of the management and supervisory boards of other companies. This can be explained by the lack of experience in board activities.

1.6 WireGraph - Entwicklung eines Lernspiels mit Prosumentenumgebung zur Förderung der fachlichen und digitalen Kompetenz von Wirtschaftsstudierenden

The sixth paper presents the educational game “WireGraph” which teaches business students on the work with knowledge graphs in the context of the Wirecard scandal.

There is great internal and external pressure for university teaching to impart digital competencies in a sustainable manner. Among other things, Big Data is mentioned as a future-oriented topic for the financial sector, which is why students of economics have to deal with technologies in the Big Data environment in order to remain competitive on the job market (Gulin et al., 2019).

Knowledge graphs are one of these technologies and working with them should be learned with the presented learning game WireGraph. In order to classify digital literacy, the European Commission proposes an appropriate reference model of digital literacy with five domains of competence and eight levels of performance, whose application to teaching-learning concepts still has research gaps (Carretero et al., 2017).

In addition, the concept of prosumer environment is proposed, which is implemented for the first time for the learning game WireGraph, to cover the competence of knowledge graph work within the competence area of digital content creation completely and across all performance levels. In doing so, students must take on the prosumer role in which they not only consume digital expertise, but also develop it. Students work through the Wirecard scandal in WireGraph using RDF and SPARQL. They then can use the scene templates to implement their own ideas and expertise in their own knowledge graph learning game (Sicilia et al., 2018).

WireGraph will be introduced for the first time in the winter semester 2021/22 and after one-year WireGraph will have been evaluated for all three phases of consumption, reflection and production. To this end, three research questions have been established and will be pursued in order to obtain initial results and thus determine the significance of the prosumer environment for future university teaching.

2 Limitations and future research avenues

2.1 Limitations

This dissertation expands the literature by examining the use of new technologies in the area of company analysis. However, some limitations should be taken into account when interpreting the results.

When analyzing the iXBRL company accounts, it should be noted that not all tags could be carefully checked due to the large amount of data. This is also not part of the analysis. Rather, the aim is to record the basic quality of the reports and the most frequent errors. In addition, the content is not checked to ensure that it was free of errors.

Data quality is always a crucial factor. This is of course also a decisive factor in the analysis and processing of UK iXBRL company accounts. Especially when creating a knowledge graph from iXBRL company accounts this is a crucial factor. Even more, as the data is derived through further data from another company house database. In addition, the relatively small sample size does not allow the discovery of a variety of patterns or relationships in the data analysis. Despite the limited dataset, one limitation we encounter is during the attempt to apply semantic reasoning on the integrated data in order to infer new relationships between the data. The inferencing process with the integrated reasoning services of Apache Jena Fuseki can not be completed for our RDF graphs possibly due to a lack of our computational capacity.

The analysis of connections of members of the management board, members of the supervisory board and auditors works sufficiently. However, it should be noted that an average business student does not have the skills to perform such an analysis. Much more IT skills are needed for this. The analyses also only took place with a limited amount of data. We have always checked the data, but since it is manually collected data, it may still contain errors.

The educational game WireGraph provides the basis for further research. At this point, the learning game has not yet been used and thus no evaluation results have been determined.

2.2 Future research avenues

The six papers composing this dissertation highlight the increasing importance and value-relevance of new technologies in company analysis.

Research on iXBRL company accounts can also be continued in many ways. For example, a different database can be used. Much more attention can also be paid to the type of companies. However, Brexit in particular offers further opportunities for future research. In general, this work provides a good basis for further research with XBRL and iXBRL reports. Especially the implementation of ESEF offers a variety of research opportunities. For example, a similar methodology can be used to analyze reports from listed companies in the EU. Thus, in a further step, a knowledge graph with ESEF reports from all European countries could also be created. This knowledge graph could then be used for auditing purposes as mentioned in the first paper.

The analysis of connections of members of the supervisory board, members of the management board and auditors can be extended in many ways. On the one hand, the data basis can be enlarged. There are a variety of ways to do this. For example, not only the DAX30 companies could be analyzed, but the entire DAX. On the other hand, it would also be possible to analyze the companies historically. Thus, it would also be interesting to see how the type and number of connections has developed historically over time. Another possibility is to compare the linkages between management boards, supervisory boards and auditors internationally.

The WireGraph educational game offers a wide range of possibilities for further implementation and research. So far, this is the first draft. In further steps, the game will be further adapted and deployed. Subsequently, the educational game can be evaluated. The focus here is on whether and how the learning game contributes to a better understanding of knowledge graphs, RDF and SPARQL. In addition, it will also be evaluated whether Wirecard's events could be communicated via the game. These findings can be used again to adapt and continuously improve the learning game.

However, the WireGraph educational game can also be further developed in other ways. Thus, the learning game is not dependent on Wirecard's history and can also be applied in other areas. In this way, the learning game can be made accessible to other groups of people in a target group-oriented manner. In addition, the learning game can also teach other competencies.

References

Bartley, J., Chen, Y.Y., & Taylor, E.Z. (2011). A comparison of XBRL filings to corporate 10-Ks – evidence from the voluntary filing program. *Accounting Horizons*, 5(2).227-245.

Business population estimates for the UK and the regions 2020. (2020). Department for Business, Energy and Industrial Strategy. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/923565/2020_Business_Population_Estimates_for_the_UK_and_regions_Statistical_Release.pdf.

Carretero, S., Vuorikari, R., & Punie, Y. (2017). The digital competence framework for citizens. *Publications Office of the European Union.*

DSW-Aufsichtsratsstudie 2018. (2018). DSW. https://www.dsw-in-fo.de/fileadmin/Redaktion/Dokumente/PDF/Presse/DSW_Pressekonferenz_Aufsichtsratsstudie_2018_-_Grafiken.pdf.

DSW-Aufsichtsratsstudie 2019. (2019). DSW. https://www.dsw-in-fo.de/fileadmin/Redaktion/Dokumente/PDF/Presse/DSW_Pressekonferenz_Aufsichtsratsstudie_2019_-_Grafiken.pdf.

DSW-Aufsichtsratsstudie 2020. (2020). DSW. https://www.dsw-in-fo.de/fileadmin/Redaktion/Dokumente/PDF/Presse/DSW_Pressekonferenz_Aufsichtsratsstudie_2020_-_en.pdf.

Gulin, D., Hladika, M., & Valenta, I. (2019). Digitalization and the Challenges for the Accounting Profession. *Proceedings of the ENTRENOVA - ENTerprise REsearchInNOVAtion Conference,* 5(1), 428–437.

Gröls, M. (2011). Die letzten Herren der "Deutschland AG"?. *Der Aufsichtsrat*, 106-107.

Hofmann, S. (2008). Handbuch Anti-Fraud-Management, Erich Schmidt Verlag, Berlin 2008.

iXBRL Tagging Features. (2019). XBRL, https://www.xbrl.org/guidance/ixbrl-tagging-features/.

Moffitt, K. C., Rozario, A. M., &Vasarhelyi, M. A. (2018). Robotic Process Automation for Auditing, Journal of Emerging Technologies in Accounting, 15(1), 1-10. 10.2308/jeta-10589.

Monk, A., Prins, M., & Rook, D. (2019). Rethinking Alternative Date in Institutional Investmet, *The Journal of Financial Data Science*, 14-31.

Oehmichen, J. (2011). Mehrfachmandate von Aufsichtsratsmitgliedern: Eine Panel-Analyze ihrer Wirkung in deutschen Unternehmen. In Lindstädt, H. (Ed.), *Schriften zu Management, Organisation und Information*, Rainer Hampp Verlag.

OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition). (2012). W3C. https://www.w3.org/TR/owl2-syntax/#Data_Properties.

Rhodes, C. (2018). *UK Business Statistics*. https://researchbriefings.files.parliament.uk/documents/SN06152/SN06152.pdf.

Robinson, I., Webber, J., & Eifrem E. (2013). *Graph databases*. O'Reilly Media, Inc.

Sicilia, M., Barriocanal, E.G., Sánchez-Alonso, S., Rózewski, P., Kieruzel, M., Lipczynski, T., Royo, C., Uras, F., & Hamill, C. (2018). Digital skills training in Higher Education: insights about the perceptions of different stakeholders. *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*. 781-787.

Singhal, A. (2012). *Introducing the Knowledge Graph: things, not strings*. Google. https://blog.google/products/search/introducing-knowledge-graph-things-not/.