IntechOpen

Data Integrity and Data Governance

Edited by B. Santhosh Kumar

Data Integrity and Data Governance

Edited by B. Santhosh Kumar

Published in London, United Kingdom

Data Integrity and Data Governance
http://dx.doi.org/10.5772/intechopen.100778
Edited by B. Santhosh Kumar

Contributors
Thanveer Jahan, Esti Stein, Yosi Ben Asher, Yury Chernov, Mantombi Maseme, Julius Olufemi Ogunleye, Nina Jeliazkova, Nikolay Kochev, Gergana Tancheva, B. Santhosh Kumar

© The Editor(s) and the Author(s) 2023

The rights of the editor(s) and the author(s) have been asserted in accordance with the Copyright, Designs and Patents Act 1988. All rights to the book as a whole are reserved by INTECHOPEN LIMITED. The book as a whole (compilation) cannot be reproduced, distributed or used for commercial or non-commercial purposes without INTECHOPEN LIMITED's written permission. Enquiries concerning the use of the book should be directed to INTECHOPEN LIMITED rights and permissions department (permissions@intechopen.com).

Violations are liable to prosecution under the governing Copyright Law.

Individual chapters of this publication are distributed under the terms of the Creative Commons Attribution 3.0 Unported License which permits commercial use, distribution and reproduction of the individual chapters, provided the original author(s) and source publication are appropriately acknowledged. If so indicated, certain images may not be included under the Creative Commons license. In such cases users will need to obtain permission from the license holder to reproduce the material. More details and guidelines concerning content reuse and adaptation can be found at http://www.intechopen.com/copyright-policy.html.

Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2023 by IntechOpen
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 5 Princes Gate Court, London, SW7 2QJ, United Kingdom

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from orders@intechopen.com

Data Integrity and Data Governance
Edited by B. Santhosh Kumar
p. cm.
Print ISBN 978-1-83968-862-1
Online ISBN 978-1-83968-863-8
eBook (PDF) ISBN 978-1-83968-885-0

We are IntechOpen, the world's leading publisher of Open Access books

Built by scientists, for scientists

6,300+
Open access books available

172,000+
International authors and editors

190M+
Downloads

156
Countries delivered to

Our authors are among the
Top 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Meet the editor

Dr. B. Santhosh Kumar is a professor at Guru Nanak Institute of Technology, Hyderabad. His research interests include data science, machine learning, blockchain technology, and data mining. He has eight patents and one copyright to his credit and has authored five books and four book chapters. He has delivered 40 guest lectures and received 14 awards from various professional bodies. He is a reviewer for several journals, including *IEEE Transactions*, *IEEE Access*, *ACM Transactions*, and others. He has served as session chair for various international conferences organized by ASDF(Association of Scientists, Developers and Faculties), Institute of Electrical and Electronics Engineers (IEEE), and Springer. He is a senior member of IEEE, appointed as an ACM Distinguished Speaker, and his profile was listed in the *World Book of Researchers* as a 2019 researcher of the year.

Contents

Preface **XI**

Section 1
Introduction **1**

Chapter 1 **3**
Introductory Chapter: Data Integrity and Data Governance
by B. Santhosh Kumar

Section 2
Data Integrity and Applications **11**

Chapter 2 **13**
Data Quality Measurement Based on Domain-Specific Information
by Yury Chernov

Chapter 3 **33**
Multiplicative Data Perturbation Using Random Rotation Method
by Thanveer Jahan

Chapter 4 **51**
FAIR Data Model for Chemical Substances: Development Challenges, Management Strategies, and Applications
by Nina Jeliazkova, Nikolay Kochev and Gergana Tancheva

Section 3
Data Governance and Applications **73**

Chapter 5 **75**
Ethical Considerations for Health Research Data Governance
by Mantombi Maseme

Chapter 6 **91**
Predictive Data Analysis Using Linear Regression and Random Forest
by Julius Olufemi Ogunleye

Chapter 7 **103**
Field Programmable Reconfigurable Mesh (FPRM)
by Esti Stein and Yosi Ben Asher

Preface

Data is one of the essential resources for an organization to perform well. We are living in an era that is highly data-driven. From decision-making processes to enhancing customer experiences, data is involved in almost all such business activities. It is the responsibility of organizations to obtain the most benefit from the numerous petabytes and exabytes of data residing in humungous databases. This is where data integrity and quality come into play. Ensuring the integrity and quality of data enriches the insights into the business operations performed. Confidentiality and safety are major concerns in this era of big data. Modifications in technologies, rapid development of the Internet and electronic trade, and the implementation of more cultured schemes for gathering, assessing, and making use of private data have made confidentiality a key focus. Data integrity is becoming more important due to the emergence of immense volumes of information being gathered and stored in computing systems. Large amounts of data acquired from diverse mediums often contain private and delicate information and thus it is of the utmost importance to safeguard this data.

Data integrity answers questions such as: When was the data created? What is its lifetime? Are the entries consistent? Data quality answers questions such as: Is the data relevant? Is the data complete? Is it unique? This book attempts to answer these questions to help individuals use data integrity and data quality to glean useful information from large volumes of data.

Section 1 includes the Introductory chapter.

Section 2, "Data Integrity and Applications", includes the following three chapters: "Data Quality Measurement Based on Domain-Specific Information", "Multiplicative Data Perturbation Using Random Rotation Method", and "FAIR Data Model for Chemical Substances: Development Challenges, Management Strategies, and Applications".

Section 3, "Data Governance and Applications", also includes three chapters: "Ethical Considerations for Health Research Data Governance", "Predictive Data Analysis Using Linear Regression and Random Forest", and "Field Programmable Reconfigurable Mesh (FPRM)".

In writing this book, I have been fortunate to be assisted by technical experts in many of the subdisciplines of data integrity and quality. First and foremost, praises and thanks to God the Almighty for his showers of blessings.

I record my indebtedness to the Chairman, Vice-Chairman, Managing Director, and Principal for their guidance and sustained encouragement for the successful completion of this book. I am profoundly grateful to my colleagues at Guru Nanak Institute of Technology for their encouragement to complete this book on time.

I am also extremely grateful to my parents for their love, prayers, and sacrifices in educating and preparing me for my future. I am very much thankful to my wife J Nandhini, my son B. S. Haarish Athithiya, and my daughter B. S. Josita Varshini for their love, understanding, prayers, and support while I worked on this project. My special thanks to my friends who inspired me to start this work. Finally, I would also like to thank the staff at IntechOpen, especially Author Service Manager Maja Bozicevic.

Finally, we would also like to thank our IntechOpen publishers for accepting and giving valuable suggestions throughout the work.

Dr. B. Santhosh Kumar
Professor,
Department of Computer Science and Engineering,
Guru Nanak Institute of Technology,
Hyderabad, Telangana, India

Section 1

Introduction

Chapter 1

Introductory Chapter: Data Integrity and Data Governance

B. Santhosh Kumar

1. Introduction

The maintenance and assurance of data accuracy and consistency throughout a system's entire lifecycle are referred to as data integrity. It is an essential component of the design, implementation, and utilization of any system that stores, processes, or retrieves data. Data integrity can be compromised in a number of ways. The word covers a large swath of ground and may signify a lot of various things depending on the precise context in which it is used, even when discussing topics that fall within the overarching category of computers. It is often used as a stand-in for the phrase data quality, but data validation is an absolute need for ensuring that data is complete and accurate. On the other hand, data integrity is the absence of any corruption in the data. Any data integrity method should ultimately strive to accomplish the same overarching goal, which is to guarantee that data are captured precisely as desired (such as a database correctly rejecting mutually exclusive possibilities). In addition, when the data is retrieved at a later time, you need to make sure that they are the same as they were when they were first recorded. In a nutshell, the goal of maintaining the integrity of data is to avoid making unauthorized modifications to the data itself.

The concept of "data integrity" should not be confused with "data security," which refers to the practice of safeguarding information against access by unauthorized individuals. The term "data integrity" refers to the quality and consistency of the data (also known as "validity") throughout the data's lifespan. At the end of the day, compromised data is of little value to businesses, and that is before we even consider the risks that come with losing sensitive data. For this reason, ensuring the data's integrity is one of the primaries focuses of the majority of business security solutions. There are a number of different ways in which the integrity of data might be undermined. When data is copied or transmitted, it must be preserved in its original form and not be changed in the intervals between updates. It is common practice to rely on error-checking techniques and validation processes in order to safeguard the authenticity of data that is copied or distributed without the aim of modifying it in any way. Another source of misunderstanding is the word "data integrity," which may mean either a condition or a process depending on how it is used. A data set is said to have data integrity if it is both valid and accurate, which are two very different things. On the other hand, data integrity as a process explains the methods that are used to assure the validity and correctness of a data set or all of the data that is stored in a database or other construct. This may be done by comparing the data against a standard. For instance, procedures for checking for errors and validating data may be referred to as "data integrity processes."

"Data governance," sometimes known as "DG" for its shortened form, refers to the process of managing the availability, usefulness, integrity, and security of the data that is stored in organizational systems. This management is carried out based on the company's internal data standards and rules, which also serve the purpose of controlling the method by which data is used. When it comes to data governance, doing it right means making sure that the data is reliable, consistent, and not utilized in any way that may be considered exploitative. A further benefit of data governance is that it assists in preventing the improper use of data. It is becoming an increasingly important factor as businesses are being forced to comply with new regulations concerning the protection of customer data and as businesses are becoming more dependent on data analytics to assist in the optimization of operations and the driving of business decisions. This is one of the reasons why it is becoming an increasingly important factor. One of the reasons why it is becoming an increasingly significant element is because of this particular aspect. A programme for the governance of data that has been established with care often consists of three distinct components: a governance team, a steering committee that serves in the function of the governing body, and a group of data stewards. Each of these components will be broken down into even more specifics in the following paragraphs. They collaborate in order to determine the rules and standards for the governance of data, as well as the techniques for its implementation and enforcement. This is largely the job of the data stewards, but it is also a part of their responsibilities. If all goes according to plan, participants in the dialog will include not only members of the IT and data management teams but also executives and other representatives of an organization's business activities [1].

Nicola Askham, an independent consultant, wrote in a blog post that she authored in January 2022 that in order for the governance programme of an organization to be successful, the organization in question must place primary emphasis on the anticipated business benefits of the programme. This was stated by Askham in a statement that was included in the post. This information was provided inside the framework of the blog post that Askham had written. This is still the case in spite of the fact that data governance is a crucial component of any all-encompassing data management strategy. During a session that took place during the 2022 Enterprise Data World Digital conference, Eric Hirschhorn, the chief data officer of The Bank of New York Mellon Corp., made a comment that was quite similar to the one that was just mentioned. The remarks made by Eric Hirschhorn may be found at this link. According to what he had to say, "excellent governance" on its own could not be considered a sufficient achievement and of itself. The final result must be an improvement in the way the organizations are handled in order to be considered successful. This extensive reference on data governance provides more clarity on what data governance is, how it operates, the benefits it brings to businesses, best practices, as well as the challenges that come with managing data. During the course of the governance process, you will not only find a discussion of numerous other pertinent technologies that may be of service to you, but you will also find an overview of data governance software. This will be included alongside the conversation. You will find hyperlinks on almost every page of this guide. These hyperlinks, when clicked, will take you to relevant sites that cover the same or similar subjects as the ones that are now being discussed. These links are scattered throughout the manual in a number of different places where they may be located [2].

2. Data integrity and data governance

In data integrity, the use of developed information evaluation techniques is required in order to investigate typically unknown, legal prototypes, and relationships in massive data sets. Mathematical prototypes, numerical procedures, and machine learning strategies might all be included in these tools. As a result, it entails the collection, organization, and storage of data, which includes evaluation and prognosis. It was possible to do this on information that was represented in quantifiable, text-based, visual, picture, or hypermedia patterns. For the purpose of doing an evaluation of the data, the apps could use certain metrics. They include things like relationship ordering or route assessments, classification, grouping, and estimations. Various businesses amass immeasurably vast amounts of information. The strategies were easily adaptable for use on traditional software and hardware platforms, which allowed for an increase in the value of already existing resources. Additionally, the strategies were adaptable enough to be combined with newly developed products and systems, which were readily available online. The repositories and information repositories are getting more and more appealing, and they are making use of the enormous number of data that needs to be evaluated effectively. It is possible that the process of data exploration in repositories entails the examination of data that is alluring, concealed, and in the normal course of events unknown from the vast repositories [3].

If the information depository has database management systems that may assist with the additional supply needs of information mining, using data integrity repositories may be more rational than using a physical subgroup of the information depository. It is recommended that a separate repository is maintained whenever it is feasible to do so. In common parlance, the phrases data integrity and data quality are often used synonymously with one another. On the other hand, they often share very few distinguishing characteristics with one another. The validation of the data and the maintenance of its unaltered state throughout its life cycle are two aspects of data integrity. On data, a wide variety of actions, including storing, retrieving, updating, and others, are carried out on a frequent basis. The procedures guarantee that the data will remain in the same format in which they were entered, regardless of the number of activities that are carried out. A few procedures, such as encrypting and backing up data, controlling who may access it, and validating it, help to keep data integrity intact. On the other hand, data is said to be of high quality if it is relevant and comprehensive as well as if it is acceptable for the purpose for which it was collected. According to the standards, data quality may be defined from three distinct vantage points, including that of the customer, that of the company, and that of the standards themselves [3].

In the case that appropriate data governance is not put into place, there is a good chance that data discrepancies that exist in a variety of systems that are spread out throughout an organization will not be handled. For instance, the systems that are used for sales, logistics, and customer support might, each in their own right, list the names of customers in a different manner from the other two systems. This might make the process of data integration more complicated and lead to problems with the integrity of the data. This, in turn, would have an effect on the precision of applications such as business intelligence, corporate reporting, and analytics. Additionally, there is a possibility that data inconsistencies will not be identified and corrected, which will have an additional negative influence on the accuracy

of business intelligence and analytics. If this occurs, there will be an additional negative influence on the accuracy of business intelligence and analytics. Inadequate data governance may also make it difficult to comply with regulatory norms, which may be very irritating for everyone engaged in the process. Businesses that are required to comply with the ever-increasing number of regulations regarding the privacy and protection of data, such as the General Data Protection Regulation of the European Union and the California Consumer Privacy Act, could find themselves in a precarious position as a direct result of this. It is generally important for an organization to establish both standard data formats and common data definitions as a part of its overall data governance programme. This is because of the interdependence between the two. In the end, increasing the quantity of data consistency, which is advantageous for the purposes of both business and compliance, is the outcome of implementing these standards and formats across all business systems [3].

3. Data governance goals and benefits

The elimination of data silos inside an organization is one of the primaries aims of data governance. Common causes of the formation of such silos include the deployment of independent transaction processing systems by distinct business units in the absence of either centralized coordination or an enterprise data architecture. The goal of the collaborative process that is data governance is to harmonize the data contained inside those systems. Participants from across all of the different business units take part in this process. A further objective of data governance is to guarantee that data is utilized appropriately. This is done for two reasons: first, to prevent the introduction of data mistakes into systems, and second, to prevent the possible abuse of sensitive information and personal data concerning consumers. This may be achieved by establishing consistent guidelines for the use of data, as well as processes that can be used to keep track of how the guidelines are being followed and ensure that they are consistently adhered to. In addition, data governance may assist in striking a balance between the practices of data gathering and the laws pertaining to privacy. Improved data quality, reduced costs associated with data management, and increased access to necessary data for data scientists, other analysts, and business users are some of the benefits that come as a result of better data governance. Other benefits include increased accuracy in analytics and enhanced regulatory compliance. By ultimately providing executives with more accurate information, data governance may ultimately aid in the improvement of company decision-making. In a perfect world, this would result in competitive advantages, more revenue, and increased profits [4].

4. Components of a data governance framework

In the context of a governance programme, the policies, rules, procedures, organizational structures, and technology that are established as part of the framework for data governance are referred to collectively as "governance framework." Additionally, it outlines things like a mission statement for the programme, its objectives, and the manner in which its success will be assessed. Furthermore, it specifies who is responsible for making decisions and who is accountable for the

different duties that will be a part of the programme. Documenting and disseminating an organization's governance structure should be done on the company's intranet for the purpose of making it immediately apparent to all parties involved how the programme will function. On the technological front, data governance software may be used to automate many parts of maintaining a governance programme. This saves a great deal of time. Even though data governance tools are not required components of the framework, having them may help with managing programmes and workflows, collaborating on the design of governance rules and process documentation, and more, in addition to supporting the building of data catalogs. Tools for master data management (MDM), data quality management, and metadata management are some examples of complementary applications that may be used with these [5].

5. Recommended procedures for the administration of data governance projects

As a result of the fact that data governance often imposes restrictions on the way in which data is handled and exploited, the technique has the potential to spark controversy inside enterprises. When it comes to information technology and data management teams, one of the most common concerns is that business users would see them as the "data police" if they take the lead on data governance initiatives. This is one of the reasons why this concern is so popular. Data governance managers who have years of experience and industry consultants both advocate that programmes be business-driven, that data owners be consulted, and that the choices on standards, policies, and procedures be made by the data governance committee. This will help eliminate pushback to governance initiatives while also promoting buy-in from businesses. Training and education on data governance is a necessary component of initiatives, particularly for the purpose of acquainting business users and data analysts with rules governing the usage of data, privacy mandates, and their responsibility for contributing to the maintenance of consistent data sets. This is particularly important for the purpose of acquainting business users and data analysts with rules governing the usage of data. In order to keep in continual communication with corporate executives, business managers, and end users about the creation of a data governance programme, a variety of outreach tools, including but not limited to reports, email newsletters, seminars, and other types of events, are necessary. This exchange of information is a precondition that must be met. A second piece of writing by Farmer presents a rundown of seven recommendations for effective data governance. Two of these seven recommended practices are training and communication, and they are both included here. Some of the others include enforcing data security and privacy standards at a location that is as close to the source system as is practically practicable, implementing suitable governance policies at every level of a business, and frequently reviewing governance policies. This proximity to the source system is important because it helps ensure that sensitive information is protected [6].

6. Data governance challenges

Due various areas of an organization sometimes have different perspectives on essential data entities, such as customers or goods, the first steps in attempts

to manage data may frequently be the most challenging. This is often the case because of the complexity of the situation. It is necessary to find a solution to these disparities as part of the process of data governance—for instance, by reaching a consensus on standard data definitions and formats. Because this may be a difficult and contentious endeavor, the committee in charge of data governance has to have a well-defined process in place for resolving disputes. The following are some more typical difficulties that businesses have while attempting to regulate their data.

Providing evidence of its worth to the company. It may be difficult to get approval, funding, and support for a data governance programme if there is no proof of the anticipated business advantages provided by the initiative up front. Askham said in a blog post that she published in January 2022 that corporate leaders want to know what is in it for them right from the beginning of a governance initiative. According to what she wrote, "if you can't answer it in a manner that they are interested in and that helps them, then they're simply not going to be interested." To demonstrate the worth of an investment to a firm on an ongoing basis demands the establishment of quantitative indicators, especially for the enhancement of data quality. This might include the amount of data inaccuracies that are fixed on a quarterly basis as well as any revenue increases or expense reductions that arise from these fixes. Other frequent metrics for measuring data quality include accuracy and error rates in data sets, in addition to associated characteristics like the completeness and consistency of the data. Learn more about the strong relationships that exist between data governance and data quality, in addition to the many types of metrics that may also be used to illustrate the efficacy of a governance programme, by reading more about these topics.

There is a considerable risk that data inconsistencies will not be managed in the several systems that are dispersed across an organization if effective data governance is not established. This is because of the widespread nature of the systems. For example, the systems that are used for sales, logistics, and customer service may, each in their own right, show the names of customers in a manner that is distinct from the manner in which the names are displayed by the other two systems. Because of this, the process of integrating the data can become more difficult, which might subsequently cause problems with the data's integrity. As a consequence of this, the accuracy of applications such as business intelligence, corporate reporting, and analytics will be impacted. Additionally, there is a possibility that data inconsistencies will not be recognized and remedied, which will have an additional negative influence on the accuracy of business intelligence and analytics. If this occurs, there will be an additional negative influence on the accuracy of business intelligence and analytics. This is because there is a chance that discrepancies in the data will not be identified and fixed, which is the reason why there is a problem. Inadequate data governance may also inhibit attempts to comply with regulatory rules, which is likely to be frustrating for everyone involved in the scenario. Companies that are required to comply with the ever-increasing number of regulations regarding the privacy and protection of data, such as the General Data Protection Regulation of the European Union and the California Consumer Privacy Act, could find themselves in a position that is problematic as a result of this. Because of this, companies that are required to comply with these regulations could find themselves in a position that is problematic. As part of their overall data governance programme, it is often vital for an organization to define common data definitions in addition to standard data formats.

7. Applications

The analytical illustration provides a trade purchasing system that makes use of the majority of the items from the year before, allowing one to make an accurate prediction of the number of products that will be required during the next time period. The authentication could check conditions such as viral, but there is a possibility that the acknowledgment and withdrawal identification may be used fraudulently. This is in contrast to the fact that viral conditions could be verified by authentication. It is put to use for a wide variety of purposes in both public and private organizations. It is common practice for businesses in the banking, insurance, medical, and buying industries to make use of data integrity in order to save costs, enhance analyses, and increase trades. Consider the insurance and banking companies that have implemented data integrity tools to assist in risk assessment and the identification of fraudulent activities. The firms could design prototypes that forecast the threats prevailing to the users in terms of credits or regarding the privileges during an accident that might be false and shall be inspected more carefully if they make use of the user-related information gathered over the course of the current period. This information was gathered over the course of the current period.

Author details

B. Santhosh Kumar
Department of Computer Science and Engineering, Guru Nanak Institute of Technology, Hyderabad, Telangana, India

*Address all correspondence to: bsanthosh.csegnit@gniindia.org

IntechOpen

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

[1] Tao X, Zhang H. Research on data security governance based on artificial intelligence technology. In: 2021 International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR). Shanghai, China; 2021. pp. 102-105. DOI: 10.1109/ICBAR55169.2021.00030

[2] Harwanto IM, Hidayanto AN. Data governance maturity assessment: A case study directorate general of corrections. In: 2022 International Conference on ICT for Smart Society (ICISS). Bandung, Indonesia; 2022. pp. 1-6. DOI: 10.1109/ICISS55894.2022.9915243

[3] Hongxun T et al. Data quality assessment for online monitoring and measuring system of power quality based on big data and data provenance theory. In: 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). China. 2018. pp. 248-252. DOI: 10.1109/ICCCBDA.2018.8386521

[4] Ladley J. Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program. Second ed United States: Academic Press Inc. 2019

[5] Nayan BR. Security and governance. In: Cloud Computing. United States: MIT Press. 2016. pp. 99-126

[6] Simon AR, Shaffer SL. Chapter 9 - Data Quality and Integrity Issues. In: The Morgan Kaufmann Series in Data Management Systems, Data Warehousing, and Business Intelligence For e-Commerce. Massachusetts, United States: Morgan Kaufmann; 2002. pp. 193-208. DOI: 10.1016/B978-155860713-2/50012-7

Section 2

Data Integrity and Applications

Chapter 2

Data Quality Measurement Based on Domain-Specific Information

Yury Chernov

Abstract

Over the past decades, the topic of data quality became extremely important in various application fields. Originally developed for data warehouses, it received a strong push with the big data concept and artificial intelligence systems. In the presented chapter, we are looking at traditional data quality dimensions, which mainly have a more technical nature. However, we concentrate mostly on the idea of defining a single data quality determinant, which does not substitute the dimensions but allows us to look at the data quality from the point of view of users and particular applications. We consider this approach, which is known as a fit-to-use indicator, in two domains. The first one is the test data for complicated multi-component software systems on the example of a stock exchange. The second domain is scientific research on the example of validation of handwriting psychology. We demonstrate how the fit-to-use determinant of data quality can be defined and formalized and what benefit to the improvement of data quality it can give.

Keywords: data quality, quality metrics, fit-to-use determinant, data warehouse, formalization, software application, test data, stock exchange, reference data, validation, handwriting psychology

1. Introduction

Over the past decades, data quality is getting more and more relevant and important both in science and practice. The topic is well known, therefore, multiple researchers and data practitioners have been intensively investigating different aspects of data quality. There are numerous publications and projects. Especially data-driven design, data-driven management, and the "big data" concept are drawing attention to the data quality issues. High quality is a prerequisite for success. The growth of BI tools (business intelligence) like Tableau or Power BI, market intelligence tools like NetBase Quid or Crunchbase Pro, A/B testing tools like Optimizely or HubSpot, etc. reflects the trend that decisions become more and more data-driven. Naturally, the requirements for data quality are permanently growing. Today, data quality is defined not only as a struggle against duplicates, outliers, missing data, corrupted text, or typos. It is a much more complicated concept.

However, the definition of data quality itself is not easy and always is a bit ambiguous. It depends on the viewpoint, aim, and context. Traditionally, data analysts define a set of data quality dimensions. Meanwhile, there are many dozens of them. Their

concepts often overlap and repeat themselves under different terms. These dimensions reflect different aspects of data quality. Most of them are well formalized and can be quantified. Quantifying is essential for planning improvement measures of data quality in different contexts.

The most popular context is natural data warehouse. Therefore, many dimensions appeared namely in this context. However, many other fields are not less influenced by data quality. The idea to develop a single quantitative indicator, which is perhaps more abstract, has been known for a long time. It seems reasonable and practical. Such a determinant is strongly domain-specific. It reflects specifics of the particular system and it can serve as a good instrument for comparison datasets.

2. Data quality dimensions

The standard approach to the definition of data quality in terms of the different aspects, which are traditionally termed dimensions, is reflected in numerous publications. Data quality dimensions were captured hierarchically in the very often referred study [1]. The study was based on a customer survey that treated data as a product. The authors defined four categories, which are still reasonable, although the study was done about 30 years ago. These categories formed the first level of the hierarchy. The data quality dimensions built the second level. Below we are following this approach and preserving the categories. However, the dimensions themselves have been modified, reflecting the development of data science and our perception of the topic. It is difficult to identify the actual source of a particular dimension. Many authors are speaking about the same dimensions often naming them differently. The review below is based mainly on several publications [2–8] that were analyzed explicitly. However, numerous additional publications, which the author studied, read, or skimmed, influenced this list as well.

2.1 Intrinsic category

2.1.1 Correctness

Data correctness or accuracy refers to the degree to which data represents real objects. In many cases to evaluate correctness, the data are compared to some reference sources. There can be different reference sources. For instance, a natural restriction to the data value (the age can be between 0 and 120 years), a certain rule (the sum of percentage values should be 100), a related database record, a calculated value, etc. When we try to formalize quality then correctness can be defined as a measure of the proximity of a data value to a referenced value, which is considered correct. If a reference source is available, an automated process of correctness evaluation can be established.

2.1.2 Validity

Validity refers to the degree to which data values comply with rules defined for the system. These can be external rules, for instance, regulation in the finance area, or internal system rules. Validity has associations with the correctness, completeness, and consistency of the data. However, data values can be valid but

not accurate, or they can be valid but not completed. Examples of non-valid data entities are a birth date, which is not in the range of valid dates, or a city, which is not in the list of cities.

2.1.3 Uniqueness/deduplication

Uniqueness means that no duplications or redundant information are overlapping across all the datasets of the system. It means that entities modeled in the system are captured and represented possibly only once within the proper component or the database segment. Uniqueness ensures that no entity exists more than once within the data. It possesses a unique key within the data set. For instance, in a master product table or person table, each product or person appears once and this entity is assigned a unique identifier. This identifier represents the product or person across the whole system. If additional instances of this product or person are created in different parts of the system, they preserve the unique identifier.

Uniqueness can be monitored statically by periodic duplicate analyses of the data or dynamically when capturing new entities. Periodic checks of the data consistency are a typical task in every data warehouse. Dynamic verifications are often built into a database as triggers and restrictions on fields. If there is a combination of numerous databases, files, and other data collection facilities, then special procedures must be developed. When some problems are detected, analysts use data cleansing and deduplication procedures to address the issue. Formal uniqueness is rather easy to ensure. More complicated are the cases when the same data are named and defined differently – formal procedures can hardly help. Artificial intelligence methods of data analysis could be useful to identify logical duplication or overlapping.

2.1.4 Integrity (referential integrity)

When we assign unique identifiers to different objects (customers, products, etc.) within our system, we simplify the management of the data. At the same time, that automatically introduces the requirement, that this object identifier is used as a foreign key within the whole data set. This is referred to as referential integrity. Rules associated with referential integrity are constraints against duplication and non-consistency.

2.1.5 Reliability/consistency

Data reliability refers to two aspects. The first aspect relates to the functioning of different data sources in the system. It should be ensured, that regardless of what source collects the particular data or where it resides, this data cannot contradict a value, which resides in a different source or is collected by a different component of the system. The second aspect relates to the closeness of the initial data value to the subsequent data value.

2.1.6 Data decay

That is the measure of the rate of negative change to data. The old values taken from different sources become outdated with time. A source can be decommissioned and a new one not applied yet. For instance, biodata, mobile numbers, and emails of persons can be not valid anymore.

2.1.7 Objectivity

It reflects the extent to which information is unbiased, unprejudiced, and impartial.

2.1.8 Reputation

It means the extent to which users regard the information in terms of source and/ or content.

2.2 Contextual category

2.2.1 Completeness

The dimension means that certain attributes should be assigned values. Completeness rules are based on the following three constrain levels:

- Mandatory attributes that require a value (for instance, a family name by a person data).
- Optional attributes, which may have a value based on some conditions (for instance, the education level of a person).
- Inapplicable attributes, which may not have a value (for instance, a maiden name for a single male).

Completeness or incompleteness can be measured through the amount of data that does not have values. The decisive is to which extent the system can perform its tasks with an uncompleted data set.

2.2.2 Data coverage

Data coverage actually reflects the second aspect of completeness, namely the completeness of records. It is the degree to which all required records in the dataset are present. Sometimes data coverage is understood as a measure of availability and comprehensiveness of data compared to the "total data universe." However, this is not practical and could hardly be quantified.

2.2.3 Amount of data

It reflects the extent to which the volume or quantity of available data is appropriate for the tasks.

2.2.4 Effectiveness or usefulness

It reflects the capability of the data set to enable users to achieve specified goals or fulfill specified tasks with the accuracy and completeness required in the context of use. Sometimes this dimension is called the relevancy or reasonability of data.

2.2.5 Efficiency

Efficiency reflects the extent to which data can quickly meet the needs of users.

2.2.6 Timeliness (currency)

Timeliness has two aspects. As data currency, it refers to the degree to which data is up-to-date and to the extent to which data are correct despite possible time-related changes.

2.2.7 Timeliness (availability)

This second aspect of timeliness refers to the extent to which data are available in the expected time frame. It can be measured as the time difference between when information is expected and when it is available.

2.2.8 Credibility

It reflects the degree to which data values are regarded as true and believable by users and data consumers.

2.2.9 Ease of manipulation

The dimension reflects the extent to which data are easy to manipulate and apply to different formats.

2.2.10 Maintainability

Maintainability is the measure of the degree to which data can be easily updated, maintained, and managed.

2.3 Representational category

2.3.1 Interpretability

The degree to which data are presented in an appropriate language, symbols, and units of measure.

2.3.2 Consistency

Consistency reflects the plausibility of data values. That is the extent to which data is presented in the same format within a record, a data file, or a database and that semantic rules are preserved all over the system. Consistency is practically the measure of the equivalence of information in various data stores and applications.

2.3.3 Conciseness

This dimension reflects how compact information is. The extent to which it is compactly represented without losing completeness.

2.3.4 Conformance/alignment

This dimension refers to whether data are stored and presented in a format that is consistent with the domain values.

2.3.5 Usability

This dimension is rather generic. It reflects the extent to which information is clear and easily used. It includes as well understandability, that is, the degree to which data have attributes that enable them to be read and interpreted by users.

2.4 Access category

2.4.1 Availability/accessibility

The dimension reflects the ease, with which data can be consulted or retrieved by users or programs.

2.4.2 Confidentiality

The degree to which disclosure of data should be restricted to authorized users. Relates to the security dimension.

2.4.3 Security

The dimension reflects the degree to which access to information is appropriately restricted.

Traceability.

It reflects to which extent data lineage is available. That is the possibility to identify the source of data and transformations they have passed.

3. The fit-for-use and domain-specific data quality determinant

Traditional dimensions of the data quality are good, since they reflect different aspects of data and are rather formal, that is, they can be in most cases automatically evaluated. However, they, first, are often derived from the data warehouse concept [9, 10] and are not always suitable in a different context. Secondly, they are good for homogeneous software systems, where they can be rather easily applied. However, they cannot be directly used for distributed heterogeneous systems, which is often the case, or for special applications, such as scientific research. Both examples we are presenting in the following text.

That is why already long ago they were speaking about the generalized fit-for-use data quality determinant [11, 12], which is close to the view of data users. That was summarized in [2]: "In general, data can be considered of high quality if the data is fit to serve a purpose in a given context." A data user can be a person, a group of people, an organization, or a software system. We consider this indicator the most important in many practical cases. Often it dominates even the formal nonconformity of a product by quality management.

Fit-for-use is a rather subjective concept. However, in the data quality context, we can provide the required formalism to make it quantitative. To enable that, we need to define a good metric. Such a metric cannot be universal—it is always context-dependent or domain-specific. However, the requirements for a data quality metric can be generic. Every good metric should answer them. In the next paragraph, we are looking at such requirements.

3.1 Requirements for Data quality metrics

In [13] authors formulated the set of requirements, which are appropriate for domain-specific data quality metrics. It includes five basic requirements:

- Normalization
- Cardinality
- Adaptivity
- Scalability (in the original publication they call it "Ability of being aggregated")
- Interpretability

Normalization should be adequate to assure that results can be interpreted and compared. That means the metric determinants should be on the same scale, which is preferably a relative one. That is important since we use data quality metrics to compare different data sets to each other and select an optimal one (our application case on testing data), to understand the trend of changes in time, or to evaluate the fitness of data for the deduced results of a scientific study (our application case on validation of handwriting analysis).

Cardinality in our context means that the metric should be highly differentiated, that is, it should ensure many possible values and not restrict itself to a rough evaluation. The sensitivity of the metrics should be good enough to capture even small differences.

Adaptability means that the metric must be easily adapted to a particular application. It should be tied to business-oriented goals. That requirement is actually the basis for the fit-to-use data quality determinant.

Scalability means that it should be possible to measure the whole system as well as its components or sub-systems. It can concern, for instance, different layers of data.

Interpretability means that the metric should be clear and simple. That means mean that a user understands the metric and that it is comprehensible and meaningful. In particular, simple metrics should be easily formalized and possibly automatically deducted from the system.

These requirements are rather technical ones. Many of the data quality dimensions mentioned above do answer them. However, sometimes and especially by fit-to-use determinants, some compromises are necessary. Metrics are needed for quantifying data quality in order to answer questions regarding the data sets and to work out the measures to improve data quality in particular domains.

4. Application case 1. Reference data of the end-to-end test system of the stock exchange

The current application case is based on the author's experience at the Swiss Stock Exchange [14]. However, the model is rather generic and it could be valid for any other financial stock exchange or other applications.

4.1 Reference data

The application case covers the quality of the reference data for the test system of the Swiss Stock Exchange. The requirements for the correctness and reliability of the system are extremely high. That is why testing plays a crucial role during the delivery of new functions into production. The controlled testing is carried out at least at four levels: component, integration, system, and end-to-end testing (unit testing is done by developers before they officially release the code). By saying "controlled" I mean that the software is delivered and built in a code control system, it has an official version number, and is installed in a controlled testing environment either automatically (DevOps) or by environment supporting stuff, that is, not by developers themselves.

Test data fully reflects the production system and consists of three parts:

- Reference (static) data remain stable—the changes can be done on a daily basis, not in real-time modus

- Configuration data, mainly technical configuration of different components, that is, IP addresses, the distribution of components over servers, timeouts, etc. They enable the system to run in the testing environment. Configuration data naturally differ from the production system. For instance, some components can share the same servers to save expensive hardware and licenses, which are not allowed in production, where reliability is the major criterion.

- Trading data are generated in real-time for testing purposes. That is done mostly using test automation scripts. However, manual intervention is as well possible.

The reference data is very important. It defines the quality of the whole testing. In the current application case, we are speaking about end-to-end testing, which is rather business than technical oriented. Therefore, the major users and customers of the test system and correspondingly end-to-end testers and business experts.

The test system reflects almost fully the real production configuration. It consists of two dozen components, which are distributed among many application servers (Linux and Windows) and multiple databases (Oracle, MS SQL, Postgres, MySQL, SQLite, and 4D). All components can be divided into three categories:

- Upstream components ensure the interface to the customers and enable the data entering.

- Trading engines, where the on-book and off-book operations take place.

- Downstream systems, such as a data warehouse or a supervision component.

The reference data in the production system is maintained partly by the system customers (banks and other trading organizations) and partly by the market supporting staff of the stock exchange. The data changes are maintained and distributed on a daily base—the reference data maintenance is not a real-time functionality. It is entered using different tools and interfaces and then is transferred into a central data repository, from which is distributed among all relevant system components. In components, the reference data can be enriched to enable more tests.

Test reference data must ensure complete test coverage. To do that, the testware is maintained in Jira and consists of the test requirements that reflect the system functional requirements, test specifications (or test plans), and test cases, which are the elements of test plans and test executions—the results of the testing for a specific project, a test cycle or a sprint. A big portion of the test cases or some steps of them are automated. Several examples of test cases regarding reference data are:

- List a new product/security of type bond.
- Update particular trading parameters of security of type share.
- Change the delisting date of security of type derivative.
- Add a new trading organization.
- Update trading access of an existing trading participant.
- Add a new market holiday.

Test cases are the major objects, which are relevant for the data quality evaluation. The testware includes both the new functionality and the regression testing, which should ensure that the current functions are not affected by new versions of software. Test data must, first, cover all existing business requirements and, secondly, additional functions that are technically possible, but are not used yet in production. They can in principle be activated later, and therefore, must be as well checked. Additionally, the data must support so-called negative test cases to test the reaction of the system to the wrongly entered data.

Reference data must be tested, since they first together with the trading data are provided to the end customers and, secondly, they are the base for the generation of the trading data (orders, trades, transaction reports, indices, etc.)

The data quality evaluation assumes that the test cases are designed properly, that is, they cover all needed functions, business-relevant cases, and configurations. A test case may reflect in this context either a business case or a certain business configuration.

4.2 Data quality determinant for reference test data

Assuming that the test case design is appropriate, we can define the following usability quality metrics.

$$q = \frac{\sum_1^n c_i b_i}{\sum_1^n c_i} \tag{1}$$

where q - data quality determinant; n - number of test cases; c_i - the weight of i-th test case; and b_i - the indicator of test case coverage by the test data (0 or 1).

The model is simple, but it is practically very useful. It fulfills the requirements mentioned above. Maybe just cardinality is fulfilled partly since the differentiability of the model is not perfect, we can get the same value of q for different weight and coverage combinations. To improve the differentiability, the determinant must be done more complex. But that will reduce the clearness and simplicity, that is, will deteriorate interpretability.

The value of the quality determinant is used to compare different test data sets to each other and to ensure the required test quality. Test data sets differ when they are applied in different test environments or at different project phases. The last is probably the most important. Therefore, if we see that the determinant in the current project (project phase) is lower than in the previous project (project phase), it is a clear requirement for additional data enrichment.

Two aspects are important: the definition of the test case weights and the evaluation (preferable automatic) of the test coverage indicator.

4.2.1 Test case weights

The test case weights are assigned by a test designer or automatically. The automation is based on the assignment of the same weight to all test cases in a certain group, typically in the same test area or test suite. Formally, weight is defined on the continuous interval from 0 to 1. Practically the following values are used: 1.0 (required), 0.75 (important), 0.5 (quite important), 0.25 (not important), 0 (not relevant). The following factors influence the weight:

- Business relevance
- Test automation
- Test case complexity
- Execution effort

Business relevance: Test cases that are more important for business should have higher weights. This aspect of test case prioritizing is covered in the publications as customer requirement-based techniques [15–18]. For instance, the issuing of a new share happens on the stock exchange four-six times a year, and at the same time, banks are listing hundreds of derivatives and structured products daily. The last case is much more important and test-relevant. Another example is that adding a new trading participant (of which there are several hundred) has a higher weight than adding a new clearing organization, which could happen once in several years. Business relevance depends on the current project. The new functions that are being introduced by the current project may have higher priority over the regression test cases, and they receive lower priority when the project is over since they become regression ones.

Test automation: Automated regression test cases have higher priority over manual ones since they check the basic functions and must be always successful. Another reason is that their execution is quicker and simpler. Therefore, they should get a weight value of 1.0.

Test case complexity: Simpler test cases should have a higher weight since the data for them could be easier maintained. Generally, a good design should lead to simple and unambiguous test cases. That is, very complex ones are in any case a bit "suspicious" and probably require re-design. They can be, for instance, broken into several simple ones.

Execution effort: Like with the complexity, test cases that require less effort should have higher priority and correspondingly higher weight by the evaluation of the test-data quality.

The initial setting of the weights requires a big effort. However, it should be generally done once and then just be maintained when new test cases are developed, or/and the system functionally is changed. That happens not very often—typically twice a year with big releases.

4.2.2 Indicator of test case coverage

The indicator of test coverage has only two values—0 or 1. When a test case cannot be executed because of the missed data then it gets the value 0. When the test case can be executed or has another problem, like a bug in the software or not implemented functionality, the indicator gets the value of 1.

The value can be assigned manually, like the weight or automatically based on the results of the test execution. For every test cycle (sprint) a test execution dashboard is defined in Jira. It includes the planned test cases and the results of the execution. The results can be passed, failed, not applicable, etc. If the result is "blocked" that means that the test case is blocked by the missing data. This information can be retrieved and used for evaluation.

4.2.3 Testware status

The current snapshot of the major end-to-end test specifications of the Swiss Stock exchange is shown in **Table 1**. The total number of test cases is 2473. Of course, that changes with new development, new projects, and corresponding versions of components.

Most test cases in the first two specifications are automated. According to the above-described logic, they get the weight value of 1.0. In the rest of the testware, some 30% of test cases have as well weight 1.0; approximately 30% - weight 0.75;

Test specification	No test cases
On-book trading	799
Off-book trading	234
Reference data	523
Post trading	137
Clearing & Settlement	186
DWH & Billing	264
Instrument submission	196
Market monitoring sanity	134

Table 1.
Testware volumes.

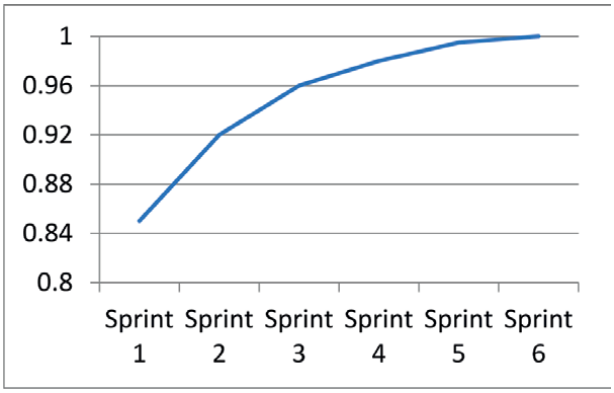


Figure 1.
Data quality dynamics.

20% - weight 0.5; and 20% - weight 0.25. This can as well differ from project to project. When a project includes software changes even in less important areas, they should be tested more thoroughly and their weight becomes higher. On the other hand, if important functionality is not affected, the test case may get a lower weight. The sum of weights in the example from **table 1** is 1929. That gives us, for instance, for the case with 20% uncovered test cases (with the weight of 1.0), what is realistic at the initial phases of a project, Q = 0.86.

The typical dynamic of the data quality (data quality determinant) along test cycles or sprints (when a project is being done with agile methodology) is shown in **Figure 1**.

5. Application case 2. Validation of handwriting psychology

The second application case relates to the area of scientific research. Data quality is not often treated formally by research experiments in psychology. Although it plays a very important role in the success and reliability of the results. An approach to quantify and model this was done by the author in the area of validation studies of handwriting analysis [19].

5.1 Handwriting analysis as a psychometric instrument

Handwriting analysis (handwriting psychology) is one of the so-called projective techniques for psychological assessment. It is based on the evaluation of a person's handwriting and deducing from it a range of personality traits. It is traditionally used for recruitment and in some specific areas where the typically used self-assessment tools are not applicable. For instance, in forensic psychology.

The technique has certain unique features and advantages over the mostly used questionnaire-based psychological tests. First, it allows the wide coverage of personal characteristics. Secondly, it excludes social desirability, which is typical for self-assessment questionnaires. However, handwriting psychology, like other projective methods, is not sufficiently validated. That often makes its usage controversial. Historically, the validation studies of handwriting analysis were based on expert procedures, involving specialists with their manual and often subjective evaluations.

In the last years, many validation studies were done with software, which does not completely substitute the experts, but rather assists them to make their evaluations more objective and reliable. One of these approaches we are discussing is below. It is based on the HSDetect system [20, 21] for handwriting analysis. The system includes statistically evaluated relations between some 800 handwriting signs and about 400 personality traits and behavior patterns.

5.2 Validation of handwriting analysis

In psychometrics, traditionally three major quality criteria are required: objectivity, reliability, and validity.

The objectivity of a psychometric test ensures that the testing person does not influence the result. In the case of handwriting analysis, the testing person is an involved expert or a computer program, which evaluates the written sample.

Reliability denotes the accuracy and precision of the procedure. The results should remain the same when the test and its evaluation are repeated under the same conditions. The typical methods for the assessment of reliability are test–retest, parallel evaluation, and split-half methods. In the context of the handwriting analysis, we consider three major components, namely, the handwriting signs, personality traits, and the relation of the signs to the traits, which we call graphometric functions. They are the objects of the quality assessment. In the traditional procedure, an expert evaluates the handwriting of the subject and interprets it in terms of personality traits, compiling a textual report. This procedure is rather subjective and that was the major objection and the root cause of the controversy. The analyzed studies were done with the computer-aided application HSDetect. This ensures objectivity and reliability [20].

Validity is the primary criterion. Objectivity and reliability are important, but they are just the prerequisites for validity. A test is then valid when it really measures what it is supposed to measure. It is always a challenge to practically define against which reference the test should be validated. Theoretically, a psychometric test should be validated against the psychological features. However, how are those obtained? In most cases, only indirectly, since a direct self-assessment is subjective and a proper expert evaluation is extremely difficult to set up. That is why a typical approach is to check the test against other psychometric tests, which are considered valid. This approach is used in statistical experiments, the data quality of which we are investigating in the current study.

The comparison between two psychometric tests is the comparison between two statistical rows – the results of the validated instrument and the reference instrument. Both tests must include the evaluation of the same subjects (involved persons). In our case, all subjects execute the reference test and provide samples of their handwriting. The handwriting is evaluated by the handwriting experts and HSDetect. Therefore, we can say that the input data for both tests are different, the output is the same - evaluated values of so-called test scales or, in other words, psychological constructs. When the results agree, we can say that the instrument under investigation (handwriting analysis) demonstrates good validity against the reference instrument. Researches very often check the agreement using correlation or another statistical method. In most of our experiments, the direct correlation does not work well and we used a special method consisting of four steps:

- Mapping of original quantitative test scale onto a simpler scale with only three values (high, medium, and low) – scale transformation.

- Assignment of all points of both the reference test and handwriting analysis to one of the three mentioned values.

- Calculation of the number of agreements (both tests have the same value) and disagreements.

- Evaluation of the statistical significance of the agreement or disagreement using the binomial distribution.

In some cases, we did use the correlation, either the product–moment correlation or the lognormal one.

5.3 Data quality determinant for the validation analysis

Data quality of a psychological test has two aspects. The first is the data of the experiment itself. Let us call it the experiment component. The second one is the distribution of subjects involved in the test along different categories – age, sex, education, profession, etc. – subject component. Both aspects are important because they both influence the meaningfulness of the test results. If we, say, make our experiments only with students, the results may be not significant for retired persons.

5.3.1 Experiment component

For the experiment component, we define the following three quality parameters:

- S - The number of subjects involved in the experiment, or in other words, the sample size.

- O – Outliers, their presence, and quantity.

- N - Normality of the row distribution.

They are briefly discussed below. By formalizing, variables S and O get a value of 1, when the quality requirement is fulfilled, or 0, when not. Variable N reflects the relation of normally distributed test scales (or test dimensions) to the total number of test scales.

Sample size: It was mentioned above that we are using the binomial check to decide the statistical consistency of the result. Typically, power analysis [22] is used to evaluate the required sample size. The standard for psychology levels of $\alpha = 0.05$ (type I error of 5%) and $\beta = 0.2$ (type II error) and the medium effect size of 0.5, results in this case in the minimal sample size = 49. Therefore, when the number of subjects is more than 48, we assume this data quality component as fulfilled, that is S = 1, otherwise S = 0.

The sample size should not be maximal big, but rather it should be optimal with adequate statistical power. It is a critical step in the design of an experiment. Involving too many participants makes a study expensive. If the study is underpowered, it is statistically inconclusive, although its results may be interesting.

Outliers: The outliers are those points of the statistical sample that are distant from other observations. This happens either due to measurement variability or due to the experiment error. Often, outliers are excluded from the data set. In this case,

they may become the subject of special analysis. The exclusion of outliers leads to a reduction in the sample size. On the other hand, that improves the experiment results. Therefore, there is always a trade-off between the result and its reliability.

In our context, there are two types of outliers. The first one means the deviation from the normal distribution of the statistical row (here is the relation to the third quality parameter N). The second type relates to the results of comparison of handwriting analysis to the psychological test. The removal of "bad" points, which mostly contribute to the disagreement, may improve the resulting evaluation. The criterion may be the proportion of improvement to the proportion of change. Say, if we remove 10% of points and that gives us 40% of improvement, we can consider the excluded points as outliers. When the improvement is 5%, the "bad" points are not outliers.

Parameter O = 1, when there are no outliers, and O = 0, if some outliers exist and were not removed.

Normality: When a random variable is normally distributed that enables many additional methods of statistical analysis, for example, correlation analysis, variance analysis, regression modeling, or ANOVA. That is why the sample must be always checked for normality. There are many methods, the most powerful of which is the Shapiro–Wilk test.

Most psychological tests have a rich normative base and they are generally normally distributed. Whether our current experiments follow the statistical population of the taken psychological test or not is not important. Therefore, we consider only the handwriting variables, which are the subject of research. In the presented model, normality in general often cannot be distinctly defined, since every test has several scales and the check is done for every particular scale and its handwriting model. Therefore, formally, the normality should be a vector and N, as mentioned above, represents the ratio of normally distributed scales to the total scales.

5.3.2 Subject component

In the experiments related to handwriting analysis, such as biodata as sex, age, handiness, education, and profession, are important, since they may influence the handwriting signs. However, in the current application case, we consider only two parameters:

- X – the sex of subjects (two values: female and male).
- A – the age of subjects (four groups are defined: below 30 years old, from 30 to 45, from 46 to 60, and above 60).

In a good experiment, both parameters should be close to uniform distribution to more or less equally represent subject categories. In this case, X and A get a value of 1. If the distribution is far from uniform, they are set to 0.

5.3.3 Data quality determinant

The determinant model is as followers:

$$q = a_S S + a_O O + a_N N + a_X X + a_A A \tag{2}$$

where a_i are corresponding weights.

		16PF-R	NEO-FFI	PVQ	EQ-i 2.0
Subjects	No	57	62	22	11
	Age 1 (<30)	15	9	6	0
	Age 2 (30–45)	17	43	8	0
	Age 3 (46–60)	15	10	6	11
	Age 4 (>60)	10	0	2	0
	Sex Male	12	27	2	3
	Sex Female	45	35	20	8
Test scales	No	16	5	10	16
	Normally distributed	12	2	7	12

Table 2.
Raw data for the estimation of quality determinant.

The defined components of data quality are not equally important. That we can solve through the standard approach—assigning different weights. Their values were defined by experts, and therefore, are rather subjective. However, that allows the comparison of different experiments. In our case, we assign the weights, so that the sum is 1.0. The experiment component gets a weight of 0.6, while the subject component is 0.4. The number of subjects is as well more important than outliers and normality. This logic results in the following weights $a_s = 0.36$, $a_O = 0.2$, $a_N = 0.21$, $a_X = 0.11$, and $a_A = 0.11$. The absolute value of the weights is not extremely important, since our aim is mostly to compare different experiments to each other.

The presented model does satisfy four requirements for the good metrics that were formulated above. Namely normalization, adaptivity, scalability, and interpretability. Only cardinality cannot be assured.

The input data for the evaluation of the quality parameters are shown in **Table 2**. We consider four studies on the validation of handwriting analysis against the following psychometric tests [20, 23]: Cattell's 16 personality factors test (revised) 16PF-R, NEO five-factor inventory by Costa & McCrae, portrait values questionnaire (PVQ) by Schwartz, and the emotional quotient inventory (EQ-i 2.0).

The data quality determinant was calculated based on model (2) and defined above weights. The results are presented in **Table 3**.

The data quality evaluation for different validation experiments demonstrates big differences. It can be a good indicator of the required experiment improvements. For instance, the removal of outliers when the sample size is big enough can be a proper way to improve the statistical power and the data quality. That may improve

Experiment	S	O	N	X	A	q
16PF-R	1	1	0.75	1	1	0.95
NEO-FFI	1	0	0.40	0	O	0.44
PVQ	0	0	0.70	1	1	0.26
EQ-I 2.0	0	1	0.75	0	0	0.37

Table 3.
Data quality determinant.

the normality of the data as well. On the other side, outliers may deliver important additional information, and, if their influence on the data quality is not that strong, they should remain in the sample.

In any case, data quality should be the important parameter for the evaluation of the reliability of the whole experiment. How to do that formally is not yet clear. That is the point for further research.

6. Conclusion

The domain-specific information is a very important factor when we are trying to define data quality. Traditional dimensions of quality reflect technical and formal aspects of the data. They are doubles useful and define the requirements for data quality. However, they are not sufficient. The real attitude of data users and the added value of the data quality is reflected in fit-for-use determinants.

In the current work, we formulate the requirements for the data quality metric and analyze two application cases with the fit-to-use determinants. They demonstrate a rather practical than theoretical approach. However, the presented results can be useful in finding ways to control data improvement.

In the presented application examples, the amount of data was small. The preparation of raw data and the estimation of the determinant value were done offline of the data. A universal data quality determinant is practically useful when it can be derived automatically from original data. The testware for the stock exchange reference data is stored in one of the test management systems (in our case, that is, Jira). The corresponding queries from the database could be easily developed and integrated with the test data management. In the second example, the required data was as well automatically derived from the experiment databases. That is a good basis for a generic system for data quality estimation. It can include the calculation engine with different models and adapters for a particular application. Their role is to retrieve the data and convert it into a generic structure.

Especially useful is a universal data quality determinant and the corresponding automatic procedure can be for artificial intelligence models. Their outcome strongly depends not only on the data quantity but as well on the quality of the training and test data. To avoid the famous GIGO (garbage in, garbage out) effect, data quality should be properly managed at all levels.

Author details

Yury Chernov
QADAS, Zurich, Switzerland

*Address all correspondence to: y.chernov@gmx.ch

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

[1] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems. 1996;**12**(4):5-33. DOI: 10.1080/07421222.1996.11518099

[2] Data VS. Quality management. In: Kunosic S, Zerem E, editors. Scientometrics Recent Advances. London: IntechOpen; 2019. pp. 1-15. DOI: 10.5772/intechopen.86819

[3] Eppler MJ. Managing Information Quality. 2nd ed. Berlin: Springer Verlag; 2003. p. 398

[4] Batini C, Scannapieco M. Data Quality: Concepts, Methodologies and Techniques. 6th ed. Berlin: Springer Verlag; 2006. p. 281

[5] Pipino LL, Lee YW, Wang RY. Data quality assessment. Communications of the ACM. 2002;**45**:211-218

[6] Redman TC. Data Quality for the Information Age. Boston: Artech House; 1996. p. 332

[7] McGilvray D. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information. Burlington: Morgan Kaufmann; 2008. p. 352

[8] Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. Communications of the ACM. 1996;**39**:86-95

[9] Kimball R, Ross M. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. 3rd ed. New York: John Wiley & Sons; 2013. p. 600

[10] Jarke M. Fundamentals of Data Warehouses. Berlin, Heidelberg: Springer Verlag; 2003. 219 p. DOI: 10.1007/978-3-662-05153-5

[11] Juran JM, Godfrey AB. Juran's Quality Handbook. 7th ed. New York: McGraw-Hill; 2016. p. 992

[12] Redman TC. Data quality management past, present, and future: Towards a management system for Data. In: Salig S, editor. Handbook of Data Quality: Research and Practice. Berlin, Heidelberg: Springer Verlag; 2013. DOI: 10.1007/978-3-642-36257-6

[13] Heinrich B, Kaiser M, Klier M. How to measure data quality? A metric-based approach. In: Proceedings of the International Conference on Information Systems (ICIS 2007); 9-12 December 2007. Montreal, Quebec, Canada: AIS; 2007

[14] Chernov Y. Test-Data quality as a success factor for end-to-end testing. An approach to formalisation and evaluation. In: Proceedings of 5th International Conference on Data Management; 24-26 July 2016. Lisbon, Portugal: Lisbon SCITEPRESS; 2016. pp. 95-101

[15] Roonqruangsuwan S, Daengdej J. A test prioritization method with practical weight factors. Journal of Software Engineering. 2010;**4**(3):193-214

[16] Zwang X, Xu B, Nie C, Shi L. An approach for optimizing test suite based on testing requirement reduction. Journal of Software. 2007;**18**:821-831

[17] Srikanth H, Williams L. On the economics of requirement-based test case prioritization. In: Proceedings of the 7th International Workshop on Economic-Driven Software Engineering Research; 15-21 May 2005. St. Louis, Missouri, USA. New York: ACM; 2005. pp. 1-3

[18] Elbaum S, Malishevsky A, Rothermel G. Test case prioritization:

A family of empirical studies. IEEE Transactions on Software Engineering. 2002;**28**:159-182

[19] Chernov Y. Data quality metrics and reliability of validation experiments for psychometric instruments. In: Proceedings of 15th European Conference on Psychological Assessment; 7-10 July 2019. Brussels, Belgium: Brussel: ECPA; 2016. p. 37

[20] Chernov Y. In: Chernov Y, Nauer MA, editors. Formal Validation of Handwriting Analysis, Handwriting Research. Validation and Quality. Berlin: Neopubli; 2018. pp. 38-69

[21] Chernov Y. Компьютерные методы анализа почерка [Computer Methods of Handwriting Analysis]. Zurich: IHS Books; 2021. p. 232

[22] Cohen J. Statistical Power Analysis for the Behavioral Sciences. New Jersey: Lawrence Erlbaum Associates; 1988. p. 567

[23] Chernov Y, Caspers C. Formalized computer-aided handwriting psychology: Validation and integration into psychological assessment. Behavioral Sciences. 2021;**10**(1):27

Chapter 3

Multiplicative Data Perturbation Using Random Rotation Method

Thanveer Jahan

Abstract

Today's applications rely on large volumes of personal data being collected and processed regularly. Many unauthorized users try to access this private data. Data perturbation methods are one among many Privacy Preserving Data Mining (PPDM) techniques. They play a key role in perturbing confidential data. The research work focuses on developing an efficient data perturbation method using multivariate dataset which can preserve privacy in a centralized environment and allow publishing data. To carry out the data perturbation on a multivariate dataset, a Multiplicative Data Perturbation (MDP) using Random Rotation method is proposed. The results revealed an efficient multiplicative data perturbation using multivariate datasets which is resilient to attacks or threats and preserves the privacy in centralized environment.

Keywords: privacy, multiplicative data perturbation, random rotation method

1. Introduction

This chapter proposes a Multiplicative Data Perturbation method. It considers multivariate datasets to perturb using a geometric data perturbation method. Then, the perturbed data will use Discrete Cosine Transformation between a pair of data values to determine Euclidean distance. This proposal is clearly elaborated in the following section.

1.1 Background

Hybrid transformations are used to maintain statistical properties of data as well as mining utilities [1–3]. The statistical properties of data are mean and variance or standard deviation without any loss of data. A feasible solution [4] is provided to optimize the data transformations by maximizing privacy of sensitive attributes. A combined technique using randomization and geometric transformation is used to protect sensitive data. A randomized technique is represented as $D = X + R$, where R is additive noise, X is original data and D is perturbed data. A geometric transformation is used as a 2D rotation data matrix represented as $D' = R(\theta) \times D$, where D is the column vector containing original co-ordinates and D' is a column vector whose co-ordinates are rotated clockwise. The above method considered only single attributes as

sensitive and rest of them as non-sensitive attributes. Data perturbation method using fuzzy logic and random rotation is proposed [5, 6].

The original data is perturbed using fuzzy based approach (M) and then random rotation perturbation is used by selecting confidential numerical attributes to get the transformed data P = M*R, where M is the dataset transformed using fuzzy based approach and R is the random dataset generated. The distorted data P is released for clustering analysis and obtained accuracy. The approach compromises in balancing privacy and accuracy. A hybrid method using SVD and Shearing based data perturbation [7] is proposed to obtain perturbed data. The approach removes the identified attributes from the dataset. These attributes are normalized using Z-score normalization to standardize to the same. Then, the dataset is perturbed using SVD transformation. Each record of the perturbed dataset is further distorted using a Shear based data Perturbation method represented as $D' = D + (Sh_D * D)$, where Sh_D is the random noise and D is the perturbed dataset obtained after SVD transformation.

The results show higher privacy is attained on hybrid methods when compared to single data perturbation methods. A hybrid technique [7, 8] based on Walsh-Hadamard Transformation (WHT) and Rotation is proposed. The Euclidean distance preserving transformation using Walsh-Hadamard (Hn) given below to generate orthogonal matrix to preserve statistical properties of the original dataset.

$$H_n = \otimes_{i=n}^{D} \quad H_2 = \frac{H_2 \otimes H_2 \cdots \otimes H_2}{n} \tag{1}$$

where H_2 is $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ is a matrix and denotes the tensor or Kronecker product. Then, Rotation transformation is applied to preserve the distance between the data points. The perturbed data preserves distance between data records and maintain accuracy using classifiers. The method is limited to numerical attributes and can be extended to categorical attributes. A hybrid approach for data transformation is proposed by Manikandan et al. [9] to sanitize data and normalize the data using min-max normalization [10]. The approach transforms original data maintaining inter-relative distance among the data. Clustering analysis shows that the numbers of clusters in original data are similar to modified data. Another approach is used to modify the original data to preserve privacy with the help of inter-relative distance on categorical data is proposed [2].

The categorical data is converted into binary data and is transformed using geometric transformation. Then the clustering algorithm is used for analysis and the results for better data utilization as well as privacy preservation. The multiplicative noise is generated using random numbers with mean as 1 and is multiplied by the original data value. A random number with a short Gaussian distribution is calculated with mean as 0 and a small variance. Geetha Mary A et al. [11] proposed a non-additive method of perturbation by randomization and data is generated based on intervals on the level of privacy specified by a user. A random number is generated that is either added or multiplied with the data to generate a random modified data. The perturbed data is classified and measures using metrics.

The condensation approach is presented by Agrawal and Yui [12] for a multidimensional perturbation technique to provide privacy for multiple columns using covariance matrix. The approach was weak in protecting data privacy. Rotation perturbation was used for privacy preserving data classification [13]. Rotation perturbations are task specific and aim to have better balance between loss of information

and loss of privacy. Multiplicative data perturbations include three types of perturbation techniques such as: Rotation Perturbation, Projection Perturbation and Geometric Data Perturbation.

A Rotation perturbation framework was adopted in privacy preserving data classification [14]. It is defined as $G(X) = RX$ where R is randomly generated rotation matrix and X is the original data. The benefit and weakness of this method is distance preservation and is prone to distance inference attacks. These attacks are addressed [15–17]. Chen et al. [14] proposed an improved version on resilience towards attacks. Oliveria et al. [17] proposed a scaling transformation along with random rotation in privacy preserving clustering.

A Random Projection perturbation is proposed [13, 18] to project a set of points from the original multidimensional space to another randomly chosen space. This resulted with an approximate model quality. A random projection matrix is used in privacy preserving data mining to enable an individual to choose their privacy levels.

An ideal data perturbation [19] aim with a balance tradeoff of minimizing information loss and privacy loss. However these are not balanced in the existing algorithms. Compared with the existing approaches in privacy preserving data mining, Geometric data Perturbation have significantly reduced these overcome [20].

A Geometric Data Perturbation is a sequence of random geometric transformation including multiplicative transformation (R), Translation Transformation (T) and Distance Perturbation (DP) [21, 22].

$$(X) = R(X) + T + DT \tag{2}$$

The approach has two unique characteristics. The first characteristic is to perturb the original data with geometric rotation, translation and identify rotation invariant classifiers as given in above. The second characteristic is to build privacy model by evaluating the privacy quality of perturbation method. The privacy model generated is used to analyze the attacks, such as, Naives and ICA-based reconstruction. The quality of data perturbation approach is determined by the quality of privacy preserved. It is the difficulty level in estimating the original data from perturbed ones such estimations are named as inference attacks. The attacks are categorized into three categories such as: Naives Inference, Reconstruction based inference and distance based inference. A statistical method based inference to estimate original data from perturbed named as Naives inference attack was proposed [23]. It is represented as O=P, where O is the observed data and P is the perturbed data. Reconstructing the data with perturbed and released information from data is presented. Reconstruction based attacks also called as Independent Component Analysis (ICA) [24, 25]. It is represented as, $O = E^{-1} P$, where E^{-1} is the estimation of released information of data and P is the perturbed data. Identifying the images and some relevant information of data using outliers to discover the perturbation is distance based attacks. It is represented as $O = E^{-1}P$, where E^{-1} is the mapping to estimate and P is the perturbed data. The higher the inference the more the original data is protected and preserved such that attacker cannot break the perturbation. The above attacks are analyzed with a privacy model with privacy guarantee [26]. It had failed to avoid outlier attack. The existing data perturbation techniques have contradiction between data privacy metric and mining utility [27, 28]. The multiplicative data perturbations will maximize the two levels i.e. data privacy and mining utility. The multiplicative data perturbation shows challenging features to improve data privacy during mining process as well as to preserve the model specific information.

In this chapter a survey is presented on privacy preserving data mining to protect confidential data. The drawbacks of the above existing data perturbation methods have made us to resolve the issues with balanced factors, such as, data privacy and data utility. The challenges in preserving privacy using multiplicative data perturbation have been given a new direction in this research study.

2. Proposed method

The proposed Multiplicative Data Perturbation (MDP) is shown at **Figure 1** as a block diagram.

The above block diagram considers the original dataset and deals with it in two stages. In the first stage, the original dataset is perturbed using geometric data perturbation. The geometric data perturbation generates a distorted dataset. This distorted dataset is further perturbed using Discrete Cosine Transformation in the second stage to finally generate a distorted dataset. The process of generating a distorted dataset using a geometric data perturbation comprises three steps. At the first step a random dataset is created using random values as in the original dataset. This random dataset is rotated counter clockwise and then multiplied with the original dataset. The resultant dataset obtained the above step is transposed in the second step, that is, Translation Transformation. This Transposed dataset is added with an additive noise in the third step to obtain a distorted dataset. This proposal is an algorithm for multiplicative data perturbation in the next section.

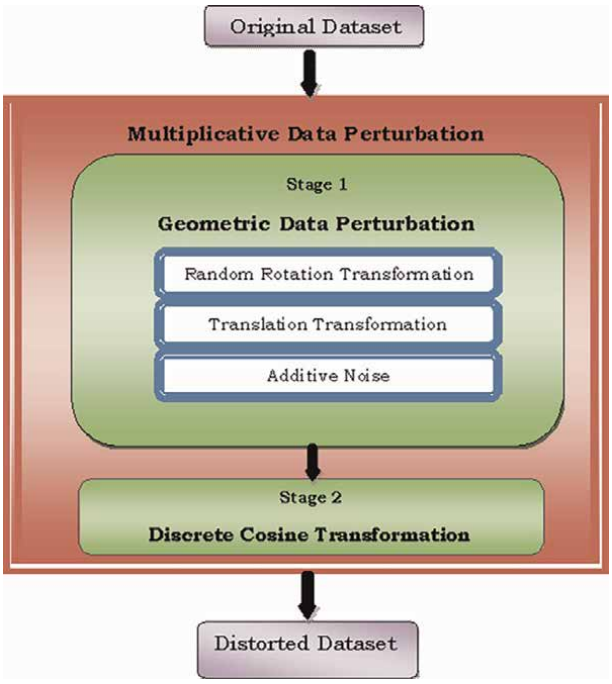


Figure 1.
Block diagram for multiplicative data perturbation using random rotation method.

3. Proposed multiplicative data perturbation using random rotation algorithm

A proposal for multiplicative data perturbation is given in this section. The pseudo code of the proposed algorithm is listed below.

Algorithm:
Input: A Data Matrix $D_{p \times q}$.
Output: A Distorted Data matrices D4, D5.
Begin.
Step 1: Create a Random data matrix R with p rows and q column and Rotate the random data matrix as $R_{q \times p}$//counter clock wise Rotation by 90°.
Step 2: Construct the data matrix $X_{p \times q}$ using $R_{q \times p}$ and $D_{p \times q}$ data matrices as:
$X_{q \times q} = R_{q \times p} * D_{p \times q}$//Multiplicative Transformation.
Step 3: Create another random data matrix $X1_{p \times q}$ with p rows, q columns with mean as 0 and standard deviation as 1.
Step 4: Construct the distorted data matrix $D4_{p \times q}$ using $X_{p \times q}$, Transpose of R and $X1_{p \times q}$ data matrices as:
$D4 = X + R^T + X1$//Geometric data Perturbation.
Step 5: Call function DCT ($D4_{p \times q}$:$D5_{p \times q}$)//Discrete cosine transformation.
Step 6: The resultant distorted data matrix $D5_{p \times q}$ is output,
End.

Function DCT ($D4_{p \times q}$:$D5_{p \times q}$)//Function for Discrete Cosine Transformation.
Input: A data matrix $D4_{p \times q}$ Output: A data matrix $D5_{p \times q}$
Begin.
Step 1: Copy the data matrix D4 to a data matrix D5//alias
Step 2: For i = 1 to q.
For k = 1 to q.
If k = 1 then

$$D4[i] = \left(1/\sqrt{i} * X_2(i) * (\cos(3.14 * (2+1)/2i))\right)$$

Else

$$D5[i] = \left(\sqrt{2}/i * X_2(i) * (\cos(3.14 * (2+1)/2i))\right)$$

End if
End For
Construct D5 data matrix and return as parameter.
End

The algorithm accepts the data matrix $D_{p \times q}$ with p rows and q columns as input. It creates a random data matrix R with p rows and q columns having random values as elements. This random data matrix R is rotated counter clockwise by 90° and then multiplied with data matrix $D_{p \times q}$. The data matrix that results is named as data matrix $X_{p \times q}$. Create another random data matrix X1 with p rows, q columns such that its mean is 0 and standard deviation is 1. Now, construct the distorted data matrix D4 adding the data matrices X, R^T and X1. This data matrix D4 is passed as a parameter to the called function DCT(). The predefined conditions are checked and data matrix D5

is updated. This data matrix D5 after completely updated is an output of the algorithm. The time complexity of the proposed MDP algorithm is found to be O(n), where n is the dimension of the dataset.

The process of updating D5 is explained with the help of an example stated below:

Example 1.1: Consider a data matrix $D_{p \times q} = \begin{bmatrix} 4 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}$ where p = 2 and q = 3.

At Step 1, create a random data matrix $R_{2 \times 3}$ as given below:

$R = \begin{bmatrix} -0.3034 & -0.7873 & -1.1471 \\ 0.2939 & 0.8884 & -1.0689 \end{bmatrix}$ and rotate R counter clockwise by 90° as given below:

$$R_{3\times 2} = \begin{bmatrix} -1.1471 & -1.0689 \\ -0.7873 & 0.8884 \\ -0.3034 & 0.2939 \end{bmatrix}$$

At step 2, construct the data matrix $X = D_{2 \times 3} * R_{3 \times 2}$ is given as below:

$$X = \begin{bmatrix} -6.4664 & -2.2049 \\ -2.2378 & 0.1134 \end{bmatrix}$$

At step 3, create another random data matrix X1 with 2 rows and 3 columns such that the mean is 0 and the standard deviation is 1.

$$X1 = \begin{bmatrix} -6.4664 & 1.4384 & -0.7549 \\ -2.9443 & 0.3252 & 1.3703 \end{bmatrix}$$

At step 4, construct the distorted data matrix $D4 = X + R^T + X1$ as given as: $D4 =$

$\begin{bmatrix} -3.1036 & -0.1362 & -1.3618 \\ -5.0820 & 2.1020 & 1.980. \end{bmatrix}$

At step 5, the function call DCT (D4:D5) where

$$DCT(k) = f(k)\sum_{k=1}^{q} D4(q)\cos[(2k+1)i\pi/2q] \quad k = 1,2\cdots q; \;\; i = 1\cdots p \qquad (3)$$

where

$$f(k) = \begin{cases} \dfrac{1}{\sqrt{q}} & k = 1 \\ \dfrac{\sqrt{2}}{q} & 2 \le k \le q \end{cases}$$

Let k = 1, q = 1, $f(k) = \frac{1}{\sqrt{q}}$., then f(1) = 1, substituting the values in the Eq. (3) $Dct(1) = 1 * -3.1036 * \cos[3 * 3.14/2] = -5.7881$

Let k = 2, q = 1, $f(2) = \sqrt{\frac{2}{q}}$, then f(2) = 1, substituting the above values in Eq. (3) $DCT(2) = 1 * -0.1362 * \cos[(2*2)*3.14/2*2] = 1.3900$

Similarly, the remaining data values of D4 are calculated to form a D5 data matrix as given below:

$$D5 = \begin{bmatrix} -5.7881 & 1.3900 & 0.4371 \\ 1.3989 & -1.5826 & -2.3630 \end{bmatrix}$$

The constructed data matrix D5 is the output.

4. Implementation

The proposed algorithm that was discussed in the previous section is implemented in MatLab. Its source code is included. The details of implementation are furnished in this section.

The implementation utilizes the built in functions available in MatLab such as load (), size(), randn(), rot90(), dct() and normrnd(). First, a load() built-in function is used to read a data into a data matrix D. The size() function is employed to retrieve the number of rows and columns. The function randn() is used to generate a random matrix R where the size is similar to data matrix D. The data matrix R is rotated using built in function available, namely rot90(). Then, to form a data matrix X, the data matrice R is multiplied by D data matrix. Next, normrnd() is called to generate a data matrix X1 having the mean as 0, the standard deviation as 1 and the size as similar to data matrix D. The distorted data matrix D4 is constructed by adding three data matrices, X1, R^T and X2. Finally, the function DCT() is employed on distorted data matrix D4 to obtain the resultant distorted data matrix D5.

5. Experimentation

The Experimentation was conducted using desktop computer system loaded with windows XP Operating system, MatLab and Tanagra data mining tool. The experimental details are elaborated in this section. The experimentation begins with the original dataset D is given as input to the proposed MDP algorithm to obtain the distorted dataset D4 and D5. Then, the original dataset D and distorted datasets D4 and D5 are uploaded into Tanagra data mining tool after appending a class attribute. These uploaded datasets are classified using classification utility available within Tanagra data mining tool. The results of classification are analyzed thereafter.

Similarly the datasets are clustered using clustering utilities available in them. The results of clustering are also analyzed and furnished at Section 6.6 under Results and Analysis. Unified column privacy metric to analyze possibility of attacks is also discussed in this section. But, their calculation is shown in section Results and Analysis. The datasets of Credit Approval, Haber-Man, Tic-Tac-toe and Diabetes are used in this experimentation. The details of Credit Approval dataset used in this experiment is furnished here and the rest of the datasets are furnished.

A Real Time Multivariate dataset, namely, Credit Approval, is downloaded from website UCI Machine Learning Repository. The details are shown at **Table 1**. Therefore the original dataset used in the experimentation is a Credit Approval dataset. It

Dataset	Size	Description
Credit Approval	690 rows & 15 columns	It consists of information of customers details concerned with credit card applications

Table 1.
Details of credit approval dataset.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
1	22.08	11.46	2	4	4	1.585	0	0	0	1	2	100	1213
0	22.67	7	2	8	4	0.165	0	0	0	0	2	160	1
0	29.58	1.75	1	4	4	1.25	0	0	0	1	2	280	1
0	21.67	11.5	1	5	3	0	1	1	11	1	2	0	1
1	20.17	8.17	2	6	4	1.96	1	1	14	0	2	60	159

Table 2.
A credit approval original dataset D.

comprises 690 rows/tuples and 15 columns/attributes including one target/class attribute.

A sample list of the original dataset D with 5 rows and 14 attributes is shown at **Table 2**.

The process in the experiment is explained as below:

First, a dataset named creditapproval.txt is loaded into X data matrix with the help of load() method. Next, the size() method on X data matrix determines the number of rows p as 690 and the number of columns q as 14. The data matrix is now named $D_{p \times q}$. Then, a built- in function randn(p, q) is used to create a random data matrix R. The random data matrix R is rotated with the help of built-in function rot90(). The data matrix X is constructed using data matrix R multiplied by data matrix D. The built in function normrnd(0,1, p, q) is used to create another random data matrix X1 with p rows, q columns, such that its mean is 0 and standard deviation is 1. Construct the distorted data matrix D4 by adding three data matrices X, R^T(transpose of R), X1. The distorted data matrix D4 is given as parameter to function DCT(D4) and it returns the final distorted data matrix D5 as output. When the above process is executed in experimentation it outputs a distorted datasets D4 and D5.

6. Results and analysis

The distorted datasets D4 and D5 together with the original dataset D, respectively are appended with a class attribute, YES or NO. The original dataset D after appending with a class attribute is shown at **Table 3**.

Similarly the distorted datasets D4 and D5 are also appended with a class attribute and furnished at section 6.6 as part of Results and Analysis. The above mentioned datasets D, D4 and D5 are uploaded into Tanagra data mining tool. First, classification utility is used on the dataset D and distorted datasets D4, D5. It divides the attributes into two categories, non-class attributes and class attribute. These two categories can be two inputs to the classifier chosen from the available ones.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	Class
1	22.08	11.46	2	4	4	1.585	0	0	0	1	2	100	1213	NO
0	22.67	7	2	8	4	0.165	0	0	0	0	2	160	1	NO
0	29.58	1.75	1	4	4	1.25	0	0	0	1	2	280	1	NO
0	21.67	11.5	1	5	3	0	1	1	11	1	2	0	1	YES
1	20.17	8.17	2	6	4	1.96	1	1	14	0	2	60	159	YES

Table 3.
A credit approval original dataset with class attribute.

Suppose we select SVM (Support Vector Machine) as classifier, then, it classifies the datasets D, D4 and D5 based on class attribute into either credit card either approved or rejected. Such results are furnished at Section 6.6 under Results and Analysis. Similarly, the experimentation is repeated with Iterative Dichotomizer 3 (ID3), (Successor of ID3) C4.5, KNN (k-Nearest Neighbor) and MLP (Multi Layer Perceptron) classifiers.

The results of those experiments are furnished at Section 6.6. A Clustering utility available in Tanagra data mining tool is used to cluster the original dataset D and distorted datasets D4 and D5. Non- class attributes are considered and given as input to k-mean clustering method. As a result, categories of clusters are formed.

A unified column metric, Root Mean Square Error (RSME) is used to evaluate inference attacks. It is calculated using Eq. (3) as given below:

$$\mathrm{RSME}(r) = \sqrt{\frac{1}{q}\sum_{i=1}^{q}(D - P)^2} \tag{4}$$

where $D = d_1, d_2 \cdots d_q$ are the original dataset values, $P = p_1, p_2 \cdots p_q$ are the perturbed dataset values and q is number of columns.

Then, privacy $(D, P) = \frac{4\sigma}{2r} = \frac{r}{2}$ (if standard deviation σ = 1). The attacks used are:

Naives inference is calculated as given in Eq. (4), where D is the original data and P = E (E is estimated or Random dataset).

Reconstruction inference is calculated as given in Eq. (4), where D is the original dataset and the Perturbed dataset

$$P = E^{-1} * P. \tag{5}$$

Distance based inference is calculated as given in Eq. (5), where D is the original dataset and P = P′ (P′ is mapped set of points of Perturbed dataset P).

The calculations of these metrics are furnished at Section 7 under Results and Analysis.

7. Results and analysis

The results obtained in the above experiment are presented in this section. The original dataset D is given as input to the proposed MDP and output distorted dataset D4 and D5 are presented below at **Table 4** and **Table 5**, respectively.

When SVM classifier is used on D, D4 and D5 datasets, the following observations are made and the same are presented at **Table 6**.

In the above **Table 4**, the first column presents the original dataset D and the distorted datasets D4 and D5. The number of tuples in the datasets considered for experimenting can be seen in the second column. The third column displays the number of training tuples classified for credit card approved as YES. The number of support vectors available is furnished in the fourth column. Fifth column reveals the error rate of SVM classifier. The computation time is tabulated at last column.

Similarly, when ID3 and C4.5 classifiers are used on D, D4 and D5 datasets the results are tabulated at **Tables 7** and **8**.

In the above **Tables 7** and **8**, the first column presents the dataset D and distorted dataset D4 and D5. The number of tuples in the datasets considered for experimenting can be seen in the second column. The third column displays the number of training tuples belonging to credit card approved as YES. A tree having number of nodes and leaves is furnished in the fourth column. Fifth column reveals the error rate of the ID3 and C4.5 classifiers, respectively. The computation time is tabulated at last column.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
7.2	15.69	14.65	5.3	18.7	9.8	2.57	−11.5	1.97	21.4	10.18	8.44	94.11	1.22
2.7	31.43	4.313	0.7	14.2	4.0	4.99	1.4	−6.69	4.77	−8.25	9.66	157.6	3.98
1.3	30.06	−13.5	17.8	3.13	22.0	−6.14	−4.23	5.28	−6.39	2.14	6.61	259.3	−2.37
9.7	26.01	9.224	−4.4	7.82	−10.3	−7.82	16.13	−10.2	1.78	13.12	−2.16	7.82	7.38
2.1	2.033	−9.22	−0.5	22.2	6.95	−0.10	2.54	−6.28	12.6	0.05	5.964	57.78	159.9

Table 4.
A credit approval distorted dataset D4.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
26.9	822.	111.6	43.7	19.6	108.	46.0	30.18	7.99	73.31	24.07	57.16	4.85	2.67
8.8	−1.6	10.90	−6.3	11.2	−4.44	6.46	6.85	−0.08	−0.08	−7.90	−5.33	−67.91	594.2
−4.8	−12.	19.15	−12.	12.4	−13.1	6.02	3.03	−10.18	−10.1	−2.99	−12.4	−229.7	−1.56
−2.3	2.86	5.801	4.62	3.55	5.80	5.89	−1.11	−11.90	−11.9	2.75	15.73	−12.28	1.87
22.9	−8.8	−11.61	−1.2	1.79	4.63	−8.72	8.00	−3.50	−3.50	2.75	−8.24	49.84	−1.27

Table 5.
A credit approval distorted dataset D5.

Dataset	Total Number of Tuples	Number of Training Tuples Classified as Approved (YES)	Number of Support Vectors	Error Rate	Computation Time (ms)
Original (D)	690	589	392	0.14	1562 ms
Distorted (D4)	690	582	621	0.446	2172 ms
Distorted (D5)	690	587	632	0.315	1969 ms

Table 6.
A credit approval dataset classified using SVM.

Dataset	Total Number of Tuples	Number of Training Tuples Classified as Approved(YES)	Tree having number of nodes and leaves	Error Rate	Computation Time(ms)
Original (D)	690	584	7 node,4 leaves	0.1464	16 ms
Distorted(D4)	690	476	3 node, 2 leaves	0.3101	31 ms
Distorted(D5)	690	580	1 node, 1 leaf	0.4464	16 ms

Table 7.
A credit approval dataset classified using ID3.

Dataset	Total Number of Tuples	Number of Training Tuples classified as Approved(YES)	Tree having number of nodes and leaves	Error Rate	Computation Time(ms)
Original (D)	690	644	67 node, 34 leaves	0.066	47 ms
Distorted(D4)	690	621	137 nodes, 69 leaves	0.101	172 ms
Distorted(D5)	690	634	157 nodes, 79 leaves	0.1246	234 ms

Table 8.
A credit approval dataset classified using C4.5.

Dataset	Total number of tuples	Number of Training Tuples Classified as Approved (YES)	Neighbors	Error Rate	Computation Time(ms)
Original (D)	690	537	5	0.2217	313 ms
Distorted(D4)	690	485	5	0.297	422 ms
Distorted(D5)	690	673	5	0.3145	391 ms

Table 9.
A credit approval dataset classified using KNN.

When KNN classifier is used on D, D4 and D5 datasets the following observations are made and presented at **Table 9**.

In the above **Table 9**, the first column presents the original dataset D and distorted datasets D4 and D5. The number of tuples in the

datasets considered for experimenting can be seen in the second column. The third column displays the number of training tuples classified as credit card approved as YES for KNN classifier. The fourth column displays the number of neighbors. The fifth column reveals the error rate of KNN classifier. The computation time is tabulated in the last column.

Similarly, the results are tabulated at **Table 10** when MLP classifier is used on D, D4 and D5 datasets.

In the above **Table 10**, the first column presents the original dataset D and the distorted datasets D4 and D5. The number of tuples in the datasets considered for experimenting can be seen in the second column. The third column displays the number of tuples classified for credit card approved as YES. The maximum number of

Dataset	Total Number of Tuples	Number of tuples Classified as Approved (YES)	Max Iteration	Train Error Rate	Computation Time(ms)
Original (D)	690	620	100	0.0924	578 ms
Distorted(D1)	690	552	100	0.168	562 ms
Distorted(D2)	690	589	100	0.347	625 ms

Table 10.
A credit approval dataset classified using MLP.

iteration for MLP classifier is furnished in the fourth column. The fifth column reveals the training error rate of KNN classifier. The computation time is tabulated in the last column. Based on the results presented above the accuracy of classification of datasets is presented at **Table 11**. The accuracy is the percentage of tuples that were correctly classified by a classifier.

The above **Table 11** presents the accuracy of the classifiers for Credit Approval, Haber Man, Tic-Tac-Toe and Diabetes datasets. The first column presents the dataset D, the distorted datasets D4 and D5. The second column presents the accuracy of classification obtained on Credit Approval dataset using SVM, ID3, C4.5, KNN and MLP classifiers. The third column presents the accuracy of classification obtained on Haber Man dataset using SVM, ID3, C4.5, KNN and MLP classifiers. The fourth column presents the accuracy of classification obtained on Tic-Tac-Toe dataset using SVM, ID3, C4.5, KNN and MLP classifiers. The fifth column presents the accuracy of classification obtained on Diabetes dataset using SVM, ID3, C4.5, KNN and MLP classifiers.

It is observed that accuracy of C4.5, KNN and MLP classifiers are better than the accuracy of the other classifiers for distorted dataset D5 compared to distorted dataset D4.

The above **Table 12** presents the comparison of accuracy. The first column presents the distorted dataset D4 and D5. The second column presents the accuracy obtained on the proposed MDP using Credit approval, Tic-Tac-Toe and diabetes datasets for SVM and KNN classifiers. The third column presents the accuracy for the existing geometric data perturbation methods using Credit approval, Tic-Tac- Toe and Diabetes datasets for SVM and KNN classifiers. It is observed that the accuracy on the datasets using our proposed MDP was found better than the accuracy of the Existing Geometric data perturbation. Moreover, their accuracy was found only on SVM and KNN classifiers for Credit Approval, Tic-Tac-Toe, and Diabetes datasets only.

The proposed MDP has given good accuracy for distorted dataset D5 compared to distorted dataset D4, whereas the literature does not show any accuracy for distorted data D5.

The results of k-means clustering are shown below at **Table 13**, when k = 2 (form two clusters).

In the above **Table 13**, the first column presents the dataset D, D4, and D5. The number of objects in the dataset considered for the experiment can be seen in the second column. The third column displays the number of objects belonging to cluster1. The fourth column reveals the number of objects belonging to cluster 2. The computational time is presented in the last column. Based on the results presented above the misclassification error rate of datasets is presented at **Table 14**.

DOI: http://dx.doi.org/10.5772/intechopen.105415

Dataset	Credit Approval					Haber Man					Tic-Tac-Toe					Diabetes				
	SVM	ID3	C4.5	KNN	MLP	SVM	ID3	C4.5	KNN	MLP	SVM	ID3	C4.5	KNN	MLP	SVM	ID3	C4.5	KNN	MLP
Original (D)	85	84	93	98	89	75	86	88	89	90	97	89	87	98	97	89	86	82	89	90
Distorted (D4)	86	70	90	88	80	68	78	79	75	76	98	73	77	96	89	80	83	79	79	85
Distorted (D5)	88	84	91	97	87	72	85	85	89	90	99	88	87	98	97	83	85	81	89	90

Table 11.
Accuracy of classifiers (%).

Dataset	Proposed Multiplicative Data Perturbation (MDP)						Existing Geometric Data Perturbation Method					
	Credit Approval		Tic-Tac-Toe		Diabetes		Credit Approval		Tic-Tac-Toe		Diabetes	
	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN
Distorted (D4)	86	88	98.7	98	80	79	86.5	82.9	98	99.5	77	73.5
Distorted (D5)	88	97	99	98.5	83.4	89	—		—		—	

Table 12.
Comparison of accuracy.

Dataset	Number of Objects	Number of Objects in Cluster 1	Number of Objects in Cluster 2	Computation time (ms)
Original (D)	690	259	431	94 ms
Distorted (D4)	690	391	299	109 ms
Distorted (D5)	690	336	354	125 ms

Table 13.
Clustering on credit approval dataset for k = 2.

Dataset	PROPOSED MULTIPLICATIVE DATA PERTURBATION (MDP)			
	Credit Approval	Haber Man	Tic-Tac- Toe	Diabetes
Distorted (D4)	0.389	0.189	0.035	0.03
Distorted (D5)	0.22	0.100	0.031	0.02

Table 14.
Comparison of misclassification error-rate.

The above **Table 14** presents the misclassification error rate. The first column presents the distorted dataset D4 and D5. The second column presents the error rate obtained on the proposed MDP using Credit Approval, Haber Man, Tic-Tac-Toe and Diabetes datasets.

In the privacy metric mentioned in Section 1.5 in Eq. 1.2, the detailed calculation of privacy quality to analyze attacks is shown below:

Consider the data matrix $D = \begin{bmatrix} 4 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}$ the corresponding distorted data matrix using the proposed MDP is given below:

$P = \begin{bmatrix} -5.7881 & 1.3900 & 0.4371 \\ 1.3989 & -1.5826 & -2.3630 \end{bmatrix}$, E is the estimated values (Random) as given below:

$E = \begin{bmatrix} -3.5441 & 1.3900 & 0.3211 \\ 0.9321 & 2.4567 & -6.7860 \end{bmatrix}$ and calculating $D'' = R^{-1}*P$ is given below

$D'' = \begin{bmatrix} -2.1461 & 0.2800 & 0.3211 \\ 1.8421 & 4.6767 & 4.6130 \end{bmatrix}$ and calculating P' is given below

Attacks	Proposed MDP Method				Existing Geometric Data Perturbation Method		
	Credit Approval	Haber Man	Tic-Tac-Toe	Diabetes	Credit Approval	Tic-Tac-Toe	Diabetes
Naives	1.743	1.129	1.564	1.512	1.345	1.234	1.456
Reconstruction	1.467	1.841	1.489	1.893	1.287	1.450	1.921
Distance	1.527	1.980	1.901	1.452	1.556	1.784	1.356

Table 15.
Analysis on attacks.

$$P' = \begin{bmatrix} -1.9261 & 0.6800 & 1.3211 \\ 3.6821 & 1.6821 & -4.5920 \end{bmatrix}$$

Then, substitute the above data matrices in eq. 1.2 to analyze the following attacks:

Naives-based Inference Attack: The RMSE is calculated by substituting the data matrices D and E. The result for RMSE r, obtained is as given below:

$$r = \sqrt{\frac{1}{3}\sum_{i=1}^{2}((D) - (E))^2} = 1.9221, \text{ Privacy } (D, P) = r/2 = 0.6796$$

Reconstruction -based Inference Attack: The RSME r is calculated by substituting the data matrices D and D″. The result r obtained is as given below:

$$r = \sqrt{\frac{1}{3}\sum_{i=1}^{2}((D) - (D''))^2} = 1.6794, \text{ Privacy } \left(D, D''\right) = r/2 = 0.839$$

Distance -based Attack: The RSME r is calculated by substituting the data matrices D and P′. The result r obtained is as given below:

$$r = \sqrt{\frac{1}{3}\sum_{i=1}^{2}((D) - (P'))^2} = 1.70261, \text{ Privacy } (D, P') = 0.851$$

Similarly the RMSE r is calculated for the original D and distorted datasets D4 and D5 and the results are furnished at **Table 15** as shown below.

In the above **Table 15**, the first column presents the Naives based, Reconstruction based and Distance -based attacks. The second column displays RMSE (Root Mean Square Error) r is calculated for the proposed MDP method on Credit Approval, Haber Man, Tic-Tac- Toe and Diabetes datasets. The third column reveals the RMSE calculated for existing hybrid methods on Credit Approval and Diabetes datasets. It is observed that the RMSE r for proposed MDP method on distance -based attack is high compared to RMSE for the existing geometric data perturbation methods. The metric for the proposed MDP shows better quality in preserving the confidential data and provides high uncertainity to reconstruct the original data.

8. Conclusion

A Multiplicative Data Perturbation algorithm by combining a Geometric Data Perturbation method and Discrete Cosine Transformation is proposed in this chapter.

The proposed MDP is successfully implemented using different multivariate datasets mentioned above.

The experiments on those datasets resulted to classify accurately and create accurate number of clusters. Based on the result analysis, it is resolved that our proposed MDP algorithm is efficient to preserve confidential data during perturbation and ensures privacy while being resilient against possible of attacks the proposed methods considered a univariate datasets ex: Terrorist. A multivariate dataset is considered and a multiplicative data perturbation (MDP) was explored to effectively perturb the data in a centralized environment. This method has resulted in perturbing the data effectively and be resilient towards attacks or threats while preserving the privacy.

The research studies can explore the privacy issues on a Big Data as a future scope of research work in the following directions:

Improving Data Analytic techniques –Gather all data, filter them out with certain constraints and use to take confident decision.

Algorithms for Data Visualization- In order to visualize the required information from a pool of random data, powerful algorithms are crucial for accurate results.

In future scope includes, research can include many various methods explore many methods. These latest methods can show various results.

Author details

Thanveer Jahan
Vaagdevi College of Engineering, India

*Address all correspondence to: tanveer_j@vaadevi.edu.in

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

[1] Li L, Zhang Q. A privacy preserving clustering technique using hybrid data transformation method. In: 2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009). Vol. 2010. Nanjing: IEEE; 2009. pp. 1502-1506. DOI: 10.1109/GSIS.2009.5408151

[2] Natarajan AM, Rajalaxmi RR, Uma N, Kirubhkar G. A hybrid transformation approach for privacy preserving clustering of categorical data. In: Innovations and Advanced Techniques in Computer and Information Sciences and Engineering. Dordrecht: Springer. 2007. pp. 403-408. DOI: 10.1007/978-1-4020-6268-1_72

[3] Selva Rathnam S, Karthikeyan T. A survey on recent algorithms for privacy preserving data mining. International Journal of Computer Science and Information Technologies. 2015;**6**(2): 1835-1840

[4] Patel A, Patel K. A hybrid approach in privacy preserving data mining. In: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). Vol. 2. Ahmedabad, Gujarat, India: IEEE; 2016. p. 3

[5] M. Naga Lakshmi and K. Sandhya Rani, "A privacy preserving clustering method based on fuzzy approach and random rotation perturbation", Publications of Problems & Application in Engineering Research-Paper, Vol. 04, Issue No. 1, pp. 174-177, 2013.

[6] Mary AG. Fuzzy–based random perturbation for real world medical datasets. International Journal of Telemedicine and clinical Practices. 2015;**1**(2):111-124. DOI: 10.1504/IJTMCP.2015.069749

[7] M. Naga Lakshmi, K Sandhya Rani," Privacy preserving hybrid data transformation based on SVD"," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, 2013, 2278-1021

[8] Jalla HR, Girija PN. An efficient algorithm for privacy preserving data mining using hybrid transformation. International Journal of Data Mining & Knowledge Management Process. 2014; **4**(4):45-53. DOI: 10.5121/ijdkp.2014.4404

[9] Manikandan G, Sairam N, Saranya C, Jayashree S. A hybrid privacy preserving approach in data mining. Middle- East Journal of Scientific Research. 2013; **15**(4):581-585. DOI: 10.5829/idosi.mejsr.2013.15.4.1.991

[10] Saranya C, Manikandan G. Study on normalization techniques for privacy preserving data mining. International Journal of Engineering and Technology (IJET). 2013;**5**(3):2701-2704

[11] Geetha Mary AN, Iyenger NSC. Non-additive random data perturbation for real world data. Procedia Technology. 2012;**4**:350-354. DOI: 10.1016/j.protcy.2012.05.053

[12] Aggarwal CC, Yu PS. A condensation approach to privacy preserving data mining. In: Proceedings of International Conference on Extending Database Technology (EDBT). Vol. 2992. Heraklion, Crete, Greece: Springer; 2004. pp. 183-199

[13] Liu K, Kargupta H, Ryan J. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering (TKDE). 2006;**18**(1):92-106

[14] Chen K, Liu L. "A Random Rotation Perturbation Based Approach to Privacy Preserving Data Classification", CC-Technical Report GIT-CC-05-12. USA: Georgia Institute of Technology; 2005

[15] Lui K, Giannella C, Kargupta H. An Attacker's view of distance preserving maps for privacy preserving data mining. In: Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases(Pkdd'06). Berlin, Heidelberg: Springer-Verlag; 2006

[16] Xu H, Guo S, Chen K. Building confidential and efficient query services in the clod with RASP data perturbation. IEEE Transactions on Knowledge and Data Engineering. 2014;**26**(2):322-335

[17] Oliveira SR, Zaiane OR. Privacy preserving clustering by data transformation. Journal of Information and Data Management (JIDM). 2010; **1**(1):37–51

[18] Guo S, Wu X. Deriving private information from arbitrarily projected data. In: Proceedings of the 11th European conference on principles and practice of knowledge Discovery in databases (PKDD07). Warsaw, Poland. 2007

[19] Balasubramaniam S, Kavitha V. A survey on data retrieval techniques in cloud computing. Journal of Convergence Information Technology. 2013;**8**(16):15-24

[20] Liu J, Yifeng XU. Privacy preserving clustering by random response method of geometric transformation. Harbin, Heilong Jiang, China: IEEE. 2010: 181-188. DOI: 10.1109/ICICSE.2009.31

[21] Balasubramaniam S, Kavitha V. Geometric data perturbation-based personal health record transactions in cloud computing. The Scientific World Journal. 2015;**2015**:927867, 1-927869. DOI: 10.1155/2015/927867

[22] Chen K, Lui L. Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining. London: Springer-Verlag Limited; 2010

[23] Hyvarinen AK, Oja E. Independent Component Analysis. New York/ Chichester/Weinheim/Brisbane/ Singapore/Toronto: Wiley-Interscience; 2001

[24] Brankovic L, Estivill-Castro V. Privacy issues in knowledge discovery and data mining. In: Proceedings of Australian Institute of Computer Ehic Conference (AICEC99). Melbourne, Victoria, Australian: Lecture Notes in Computer Science. 1999;**4213**:297-308. DOI:10.1007/11871637_30

[25] Liu K, Giannella C, Kargupta H. An Attacker's view of distance preserving maps for privacy preserving data mining. In: European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). Berlin, Germany; 2006

[26] Li L, Zhang Q. A privacy preserving clustering technique using hybrid data transformation method. In: Grey Systems and Intelligent Services, 2009 GSIS 2009, IEEE International Conference. Nanjing, China: IEEE; 2010. DOI: 10.1109/GSIS.2009.5408151, 08

[27] Rajesh N, Sujatha K, Kumar AALS. Survey on privacy preserving data mining techniques using recent algorithms. International Journal of Computer Applications Foundation of Computer Science (FCS). 2016;**133**(7):30-33

[28] Patel L, Gupta R. A survey of perturbation technique for privacy-preserving of data. International Journal of Emerging Technology and Advanced Engineering Website. 2013;**3**(6):162-166

Chapter 4

FAIR Data Model for Chemical Substances: Development Challenges, Management Strategies, and Applications

Nina Jeliazkova, Nikolay Kochev and Gergana Tancheva

Abstract

Data models for representation of chemicals are at the core of cheminformatics processing workflows. The standard triple, (structure, properties, and descriptors), traditionally formalizes a molecule and has been the dominant paradigm for several decades. While this approach is useful and widely adopted from academia, the regulatory bodies and industry have complex use cases and impose the concept of chemical substances applied for multicomponent, advanced, and nanomaterials. Chemical substance data model is an extension of the molecule representation and takes into account the practical aspects of chemical data management, emerging research challenges and discussions within academia, industry, and regulators. The substance paradigm must handle a composition of multiple components. Mandatory metadata is packed together with the experimental and theoretical data. Data model elucidation poses challenges regarding metadata, ontology utilization, and adoption of FAIR principles. We illustrate the adoption of these good practices by means of the Ambit/eNanoMapper data model, which is applied for chemical substances originating from ECHA REACH dossiers and for largest nanosafety database in Europe. The Ambit/eNanoMapper model allows development of tools for data curation, FAIRification of large collections of nanosafety data, ontology annotation, data conversion to standards such as JSON, RDF, and HDF5, and emerging linear notations for chemical substances.

Keywords: FAIR, database, data model, chemical substance, nanomaterial, structure, molecular descriptors, linear notation, ontology

1. Introduction

Since the emergence of the term cheminformatics within the context of pharmaceutical industry activities around the end of the twentieth century, an adequate chemical structure representation has been essential for the efficient application of cheminformatics methodologies [1]. The chemical structure is at the core of various cheminformatics activities: molecular property prediction via Quantitative Structure-Property Relationships/Quantitative Structure-Activity

Relationships (QSPR/QSAR), searching new biologically active compounds, lead optimization, virtual screening, combinatorial chemistry, etc. The centrality of molecular structure gives the primary flavor that distinguishes these activities from the classical chemometrics approaches [2], focused on data mining of the analytical and experimental results in order to extract useful information for the chemical objects study (e.g. the popular structure elucidation task). The chemometrics techniques from the 70s were transferred, adapted, and further developed within the field of "mathematical chemistry" with a strong focus on graph theory applications for molecule structures representation in the 80s and in 90s, and together with the 3D structure information focus and movement toward big data, resulted in the birth of modern cheminformatics. The main motto "from data to knowledge" summarizes the data workflow from studying chemical objects toward gaining/formalizing chemical information and generation of chemical knowledge as models, classifiers, etc. An adequate representation of the structures is required for all stages of the data management workflow. The chemical object representation development is a dynamic process, which is strongly influenced by the practical needs of the industry and lately, regulatory bodies. The novel deep learning technologies are changing the ways the structure information is used (e.g. linear notations can be directly read by the artificial neural networks as well as vector representations of the structures generated). The chemical substance model is a logical extension of the traditional molecule representation and takes into account practical aspects of chemical data management and new emerging research challenges. Finally, the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [3] were widely popularized and strongly encouraged as a needed background for efficient ongoing interconnections and activities within the: academia, industry, and regulators. On the other hand, the substance paradigm is based on a more complex approach toward representation of the chemical objects and must handle multiple material compositions, enriched with mandatory metadata and corresponding ontology annotations in order to comply with FAIR principles.

In the following sections, the reader will be led into a journey, starting from the classical molecular data model, based on chemical structure, and going through complex representations of chemical substances, industry use cases, and nanomaterials (NMs). The logical evolution of the data model elucidation will be demonstrated within the context of various challenges. The importance of metadata will be discussed as well as the adoption of FAIR principles. The good practices will be exemplified by the Ambit/eNanoMapper data model and real chemical substances from ECHA REACH dossiers. This chapter also discusses the FAIRification of large collections of data and the importance of standard data formats and emerging linear notations for chemical substances.

2. Classical cheminformatics paradigm for molecular data: structure, properties, and descriptors

The chemoinformatics is a vast interdisciplinary field with a large inheritance from the data mining, graph theory, and mathematical chemistry, enriched with modern methods for big data and artificial intelligence approaches. A common denominator of this methodological variety is the focus on the chemical structure. The centrality of chemical structure is also evident in other domains, strongly related to the cheminformatics, such as reaction informatics, bioinformatics (e.g. proteomics and metabolomics), toxicogenomics, etc. In QSPR/QSAR analysis, physicochemical

properties and biological activities are considered as functions of the molecular structure, i.e. P = f(S) or A = f(S). Also, equation reversal is observed for the chemometrics' structure elucidation task: $S = f^{-1}(P)$, e.g. structure is obtained out of the spectral data (spectrum is the property vector, P, in this case). The representation of the chemicals is the starting point for any of these activities. The molecular structure is the principle "model" that encompasses most important bits of the current chemical knowledge, used for further data processing and modeling.

The hierarchy of basic chemical objects' representations is shown in **Figure 1**. It starts from the smallest chemical objects, atoms, and bonds, which are the building blocks for the chemical structure. The connection table (CT) encodes the chemical graph and is the most widely used approach for structure information representation on a topological level. 2D coordinates and 3D coordinates together with the CT fully describe a chemical structure. Traditionally, the transition from structure to property is helped by an intermediate layer of descriptors, D, i.e. the first step is $D = f_1(S)$ and then $P = f_2(D)$.

The classical and widely adopted data model of a molecule representation is defined as a triple of the type (S-structure, D-descriptors, and P-properties), as illustrated in **Figure 2**. Different structure representations are systemized in several levels: 0D/1D – constitution, 2D – topology, 3D – geometry, 4D – conformation, and the QM (quantum mechanics) level with detailed electronic structure information. The intermediate layer of molecular descriptors is derived computationally or experimentally and represents useful information for the molecule. Structural descriptors are an important subset of descriptors, used as the principle interface between structure and properties. The structure is reduced to a simpler representation, namely a point in n-dimensional vector space. Variety of cheminformatics tasks, such as searching, classification, virtual screening, clustering, and measuring distance between the objects, can be performed in terms of points in the chemical space of so called "patterns." Traditionally, the chemical patterns are considered more user-friendly to the classical machine learning methods than the original chemical objects.

Figure 3 shows various structure representations for the molecule of benzene: connection table, 2D and 3D coordinates (with corresponding graphical model), linear notations – SMILES, InChI and SLN, distance matrix as a topological descriptor and registry numbers CAS N, (EC) Number, and PubChem CID. Also, **Figure 2** exemplifies different descriptors: constitutional (N_A, N_{DB}), topological (Wiener index and kappa1 index), and geometrical (eccentricity and radius of gyrations) plus the third data layer with molecular properties: LogP, RI, BP, etc.

The majority of chemical database implementations are based on the classical structure paradigm – the molecule triad (S, D, P). This paradigm has been used for several decades, and even nowadays it is the predominant base layer for the public

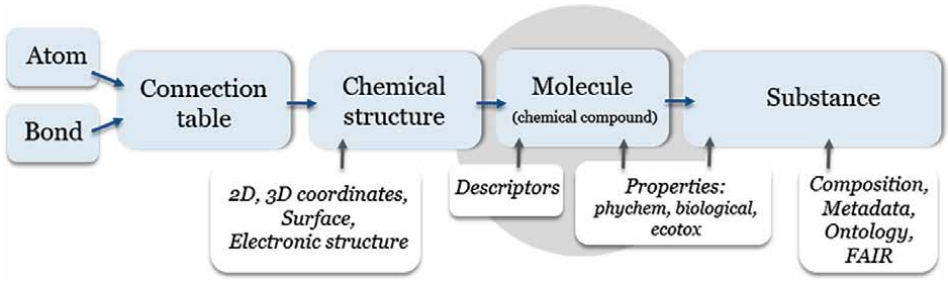


Figure 1.
Hierarchy of chemical objects: From primitive/small objects (left) to larger and complex objects (right).

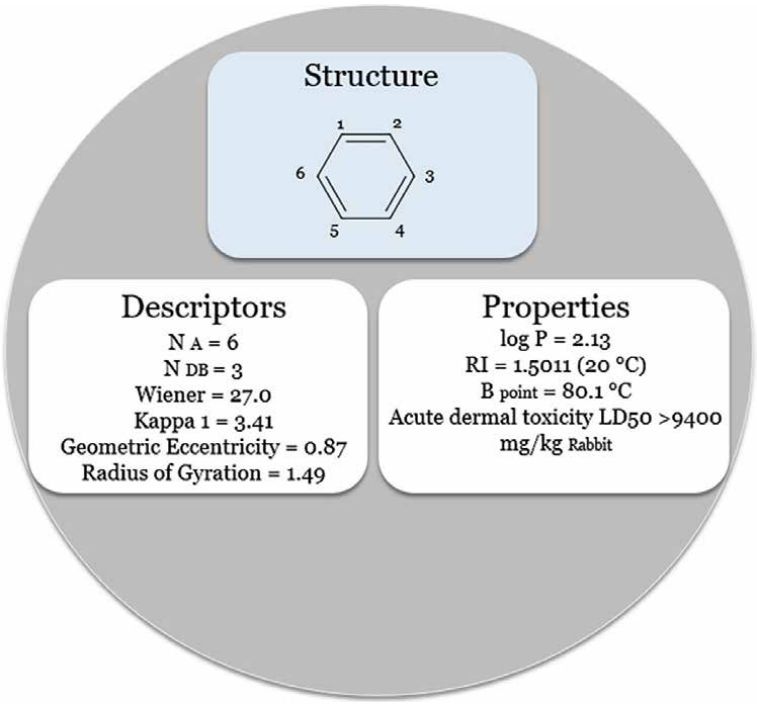
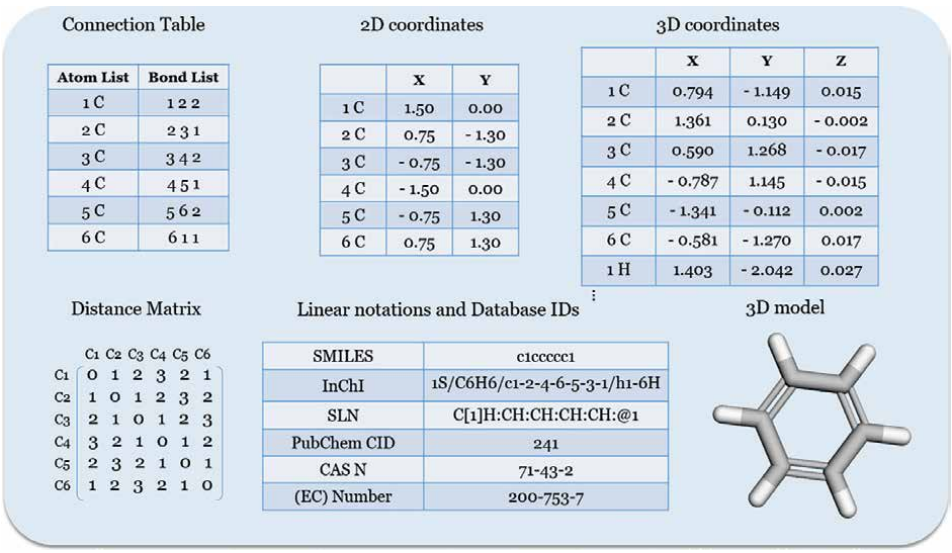


Figure 2.
Classical triad model of a molecule: (structure, descriptors, and properties).



Connection Table

Atom List	Bond List
1 C	1 2 2
2 C	2 3 1
3 C	3 4 2
4 C	4 5 1
5 C	5 6 2
6 C	6 1 1

2D coordinates

	X	Y
1 C	1.50	0.00
2 C	0.75	- 1.30
3 C	- 0.75	- 1.30
4 C	- 1.50	0.00
5 C	- 0.75	1.30
6 C	0.75	1.30

3D coordinates

	X	Y	Z
1 C	0.794	- 1.149	0.015
2 C	1.361	0.130	- 0.002
3 C	0.590	1.268	- 0.017
4 C	- 0.787	1.145	- 0.015
5 C	- 1.341	- 0.112	0.002
6 C	- 0.581	- 1.270	0.017
1 H	1.403	- 2.042	0.027

Distance Matrix

	C1	C2	C3	C4	C5	C6
C1	0	1	2	3	2	1
C2	1	0	1	2	3	2
C3	2	1	0	1	2	3
C4	3	2	1	0	1	2
C5	2	3	2	1	0	1
C6	1	2	3	2	1	0

Linear notations and Database IDs

SMILES	c1cccccc1
InChI	1S/C6H6/c1-2-4-6-5-3-1/h1-6H
SLN	C[1]H:CH:CH:CH:CH:@1
PubChem CID	241
CAS N	71-43-2
(EC) Number	200-753-7

3D model

Figure 3.
Various structure representations, descriptors, linear notations, and registry (database) indices for the molecule of benzene.

chemical databases. Naturally, the cheminformatics community and academic circles feel quite comfortable within the triad model. It has been like a "protecting bubble" and proved its usefulness as a "ground zero" for the chemical information workflow.

However, staying in the limits of the classical (S, D, P) model may hinder or isolate the cheminformatics field evolution. The "conveniences" and simplicity of the (S, D, P) model may prevent the establishing of efficient interconnections between the cheminformatics field and other scientific areas, especially in the context of industry and regulators. In the following sections, we describe the further development of the classical triad model into the paradigm of chemical substances (see the last element from the chemical objects chain in **Figure 1**).

3. Data models for chemical substances

The chemical structure describes a well-defined molecule. Unlike chemical structures, real chemical objects or industrially manufactured ones are not pure substances. Such substances are composed of several components; hence, they cannot be associated with a single unique structure. The regulatory authorities typically need information on chemicals as produced by industry. Another data gap emerges from the lack of tools to consider metadata about the performed experiments and measurements in cheminformatics use cases, e.g. QSAR model building, while such metadata is crucial for the toxicologists and regulators. The substance have to be represented as the entirety of the components with their roles and relations, include rich metadata to enable unambiguous description of experimental results from many biological assays, physicochemical characterizations, exposure, and environmental fate tracking. The challenges are increasing with representation of nanomaterials and advanced materials. Having a consensus on the chemical substance definition is a challenge also due to the discrepancies between the approaches of various regulatory institutions.

According to the International Union of Pure and Applied Chemistry (IUPAC) definition [4], a substance "*is matter of constant composition best characterized by the entities (molecules, formula units, atoms) it is composed of. Physical properties such as density, refractive index, electric conductivity, melting point etc. characterize the chemical substance*." Under Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), the concept of substance is clearly described [5]: "*A substance is a chemical element and its compounds in the natural state or the result of a manufacturing process. In a manufacturing process, a chemical reaction is usually needed to form a substance*." Under REACH, a chemical substance is composed of three types of components: constituents, impurities, and additives. Chemical substances can be mono-constituent (one main constituent is present to at least 80% (w/w)), multi-constituent (more than one main constituent is present in a concentration between 10% and 80% (w/w)), or UVCB (Substance of Unknown or Variable composition, Complex reaction products or Biological materials). The REACH definition of a substance encompasses all forms of substances and materials on the market, including nanomaterials.

The Government of Canada [6] defines a chemical substance as: "*Elements or compounds that are deliberately created, produced as a by-product of other processes or occurring naturally in the environment.*" The Canadian Environmental Protection Act (CEPA) also requires notification for new substances put in two lists: the Domestic Substances List (DSL) and the Non-DSL (NDSL). Toxic Substances Control Act (TSCA) [7] requires the United States Environmental Protection Agency (US EPA) to compile, keep current, and publish a list of each chemical substance that is manufactured or processed, including imports, in the United States for uses under TSCA. TSCA defines a "chemical substance" *as any organic or inorganic substance of a particular molecular identity, including any combination of these substances occurring in whole or in part as a result of a*

chemical reaction or occurring in nature, and any element or uncombined radical. The Japanese Act on the Evaluation of Chemical Substances and Regulation of Their Manufacture is performed under Chemical Substance Control Law (CSCL) [8].

The underlying data model is of crucial importance for the efficiency of any cheminformatics, nanoinformatics, and bioinformatics workflow. Specifically, nanomaterial (NM) representations are the primary subject of the new and rapidly evolving field of nanoinformatics. According to ISO TS 80004–1:2015, definition of a nanomaterial is: "*a material with any external dimension in the nanoscale approximately 1 nm to 100 nm and/or having internal structure or surface structure in the nanoscale*" [9]. The European Commission [10] definition of a nanomaterial is: "*A natural, incidental or manufactured material containing particles, in an unbound state or as an aggregate or as an agglomerate and where, for 50 % or more of the particles in the number size distribution, one or more external dimensions is in the size range 1 nm-100 nm. In specific cases and where warranted by concerns for the environment, health, safety, or competitiveness, the number size distribution threshold of 50 % may be replaced by a threshold between 1 and 50%.*" The substance definition in the European Union regulation REACH [5] and in the Classification, Labelling and Packaging (CLP) Regulation includes all forms of substances and materials on the market, including NMs, i.e. NM is treated as a particular case of a chemical substance.

There are several major data models highlighting the path for storing chemical substances in a database. IUCLID [11], the primary software for preparation and submitting REACH dossiers, stores and maintains data on the hazardous properties of chemical substances and mixtures, as well as their use and associated exposure levels. This is also the first system that fully implements the OECD harmonized templates (HT) [12] on the base of OECD guides of testing and agreed standards. The BioAssay Ontology (BAO) [13] provides a foundation for standardizing assay descriptions and endpoints with capabilities enabling the retrieval of data, relevant to a query. This is the first ontology to describe this domain, and certainly the first time that bioassay and HTS (high throughput screening) data have been represented using expressive description logic [14].

CODATA, the International Council for Science: Committee on Data for Science and Technology (www.codata.org), and VAMAS, an international pre-standardization organization, concerned with materials test methods (www.vamas.org), jointly foster the development of a uniform description system for NMs to address the diversity and complexity of nanomaterials. CODATA [15] encourages the interoperability and the usability of such data using a framework with four basic information categories General Identifiers, Characterization, Production and Specification and numerous subcategories and descriptors for detailed information. Most of the terms and concepts used in the descriptive system are easily understandable for people from different directions, as it is expected to be used by different groups of users for research reports, NM identification in regulations and standards, specifying NMs in commercial transactions, etc. [16].

ISA-TAB [17] defines three basic layers for sharing metadata, related to experiments: Investigation, Study and Assay, and the actual experimental data is stored on a separated forth layer and referenced by the ISA data [18]. Additional configuration settings and ontology annotations could be considered as additional layers to this complex multi-layered approach. The ISA model is non-standardized and user-defined and can include image files, spreadsheets, and protocol documents, forwarded to appropriate fields in the Study file table. The basic approach to present chemical compounds in ISA-TAB is an ontological record, which usually points to a single chemical structure. ISA model can be

serialized via ISA-TAB [19] format as multiple spreadsheet files or ISA-JSON [20] – data is stored in more convenient fashion as JSON (JavaScript Object Notation [21]) files.

Although the technical approaches and the use case scenarios of the four data models differ, a unifying logic could be traced. The need of generally "larger" chemical data object is not seen only in ECHA's REACH dossiers but also in all regulatory platforms (e.g. CEPA, TSCA, etc.) as well as such courses of action could be observed in public chemical databases evolution (e.g. PubChem has the notion of chemical substance). The foundation of a more sophisticated data model for substances is laid with three principal pieces of information: (1) identification, (2) material/substance description and composition, and (3) measurements records. This is practically illustrated in **Figure 4** for the substance of "benzene." The substance data model obviously includes a collection of standard triples:

Structures = $\{(S_k, D_k, P_k) \mid k = 1,2,\ldots,m\}$,

However, a collection of new objects of the type "chemical substance" is needed to encompass the three principle levels. An identification layer may include identifiers and names. The challenge of unique substance identifiers and names is discussed in the last section of the book chapter.

A dynamic approach of material description is needed in at least two dimensions. Apart from multiple components, the industry and regulators, also, need to handle multiple compositions of the same substance, e.g. there may be different manufacturing processes for the same products. The latter is demonstrated with two different compositions of the "benzene" substance, as shown in **Figure 4**. The first one contains three components: benzene as main constituent, toluene as impurity, and some nonaromatic hydrocarbons. Toluene, which is an impurity in composition 1, is included in the second composition as well, but with a different role – it is a constituent of the "benzene" substance. The data model requires new data entities like:

Substance = {.
$\{name_1, name_2,\ldots, id_1, id_2,\ldots\}$;

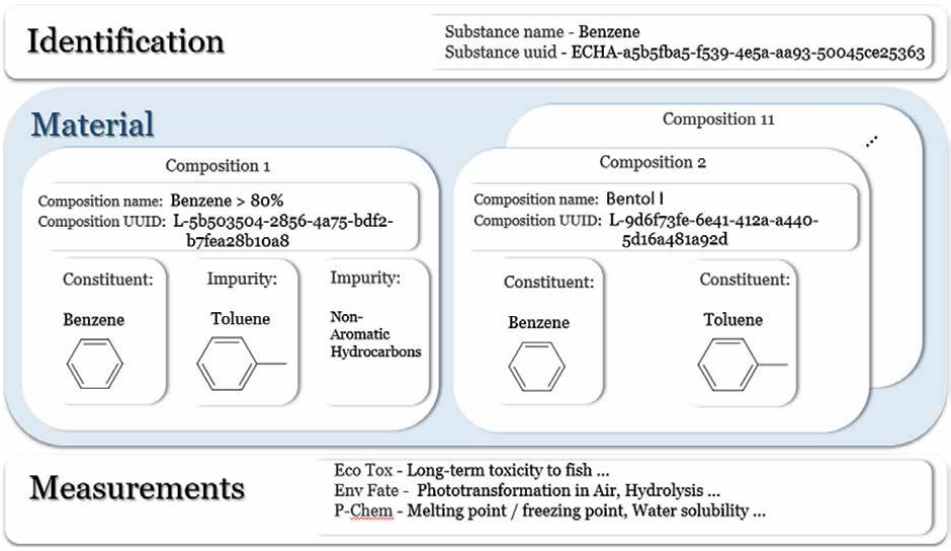


Figure 4.
Substance of benzene with several different compositions and information grouped in three layers: Identification, material, and measurements (example is taken from the public records of the ECHA's dossiers and is also accessible via ambit-LRI database web interface).

{composition$_1$, composition$_2$,...};
{measurement$_1$, measurement$_2$,...};
}
and
Composition = {.
{(S_1, D_1, P_1), (S_2, D_2, P_2),...};
{component relations};
{component concentrations};
}

In **Figure 4**, the term "benzene" is used for naming two different types of objects. There is a chemical structure of the benzene molecule which is the main constituent of the "benzene" substance. Hence, clear communication requires proper context in terms of data object types. On the other hand, the molecule of benzene could participate in other chemical substances with different roles. As it is illustrated in **Figure 5**, benzene molecule is an impurity component.

Also, the composition data entity should not be mistaken with the substance entity as well as the structure identifiers (e.g. benzene molecule CAS Number, 71–43–2 and InChI = 1S/C6H6/c1–2–4-6-5-3-1/h1-6H) should not be mistaken with the substance identifiers. The latter is a subtle error but is a common mismatch due to a long-term dominance of the structure-centered thinking. For example, in the nanoinformatics field, CAS number is wrongly associated with the whole nanomaterial instead of with a particular NM component. Also in **Figures 4** and **5**, identifiers of the substance compositions are shown as well. The complicated relationships between the three types of entities: structures, substances, and compositions require identifiers for all entity types. The shown examples utilize internal hash-based identifiers, uniquely generated by the Ambit-LRI database system. The principle difference between

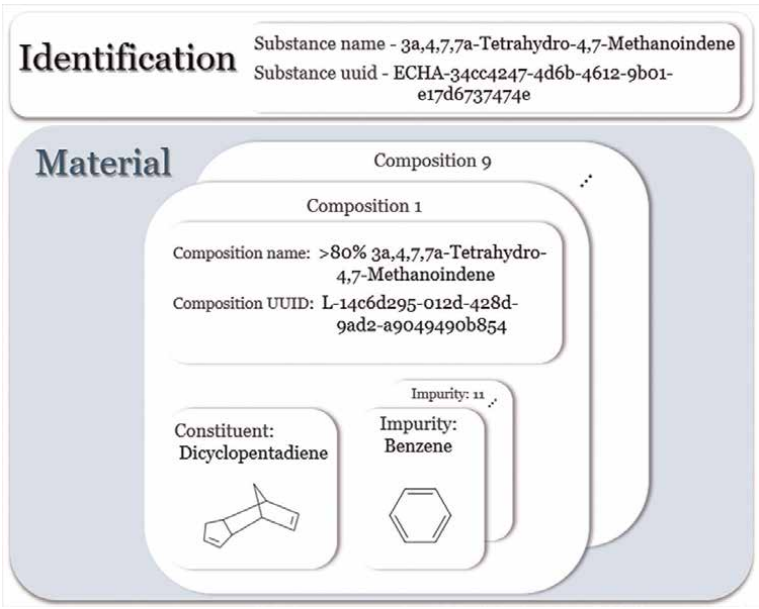


Figure 5.
The substance of "3a,4,7,7a-tetrahydro-4,7-methanoindene" with a benzene molecule as an impurity component within the first composition (example is taken from the public records of the ECHA's dossiers and is also accessible via ambit-LRI database web interface).

structure and substance also dictates different approaches for database searching for these types of objects. Structure collections can be searched with the well-known cheminformatics methods (e.g. identity search, similarity, and substructure search), but the resulting hit list with structures should be logically related with the other types of data entities, i.e. substances and compositions. For instance, benzene structure is a component, playing different roles within about 200 different substances from the public ECHA's dossiers. Also, within the context of experimental data handling, there is a difference between the properties of the "entire" chemical substance, stored on the Measurements layer (see **Figure 4**) and the "nominal" properties of the component, as they are treated in the standard triad model.

4. FAIR principles

A chemical substance database is expected to facilitate analysis of chemical and physical properties, biological analyses, and human and environmental impacts, particularly in the context of safety and risk assessment. Integration of data from multiple sources (e.g. for the needs of read across) is only possible if original measurements are combined with rich metadata and obey a set of well-established good practices for data management. In 2016, Scientific data [3] published "FAIR Principles for Scientific Data Management." The authors provide guidance for improving the discoverability, accessibility, interoperability, and reuse of data popularized as FAIR (Findable, Accessible, Interoperable and Reusable). The principles (see **Figure 6**) emphasize machine capability (i.e. the ability of computing systems to find, access, interact with, and reuse data with no or minimal human intervention), since humans increasingly rely on computational support for data processing as a result of the increase of the volume, complexity, and speed of data creation.

The four foundational principles guide data producers and publishers in how to increase the value of modern scholarly digital publishing. The application of FAIR principles is, also, to algorithms, tools, and workflows used for data generation.

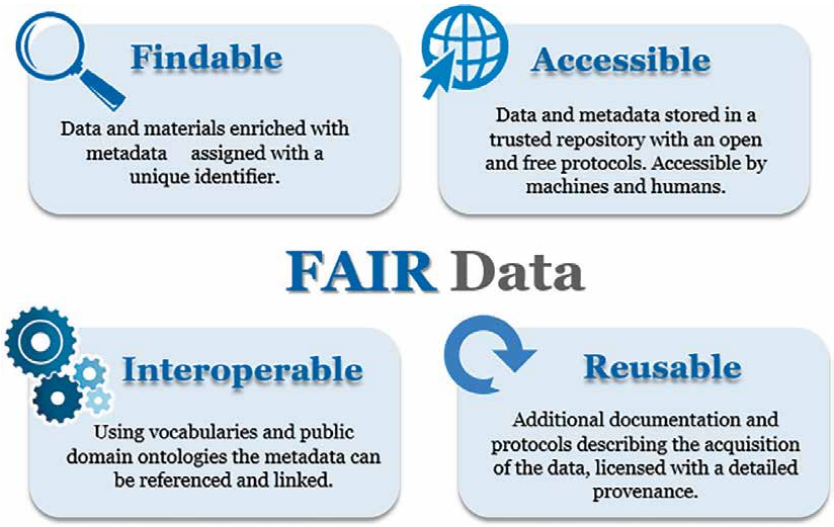


Figure 6.
FIAR principles: findable, accessible, interoperable, and reusable.

GO-FAIR initiative (https://www.go-fair.org/) gained much popularity in the last few years and strongly endorses deployment of as much as possible FAIR data resources. On their dedicated site, GO-FAIR recommends a workflow of seven basic stages for transforming a non-FAIR data resource into a FAIR one (**Figure 7**). Rich and descriptive metadata, used by machines, is a key tool to evaluate and answer the questions being asked about the data. FAIR principles allow experimental data to be used beyond their origin to solve scientific problems, fill in missing data, reuse data in applications, do modeling, and provide tools for other scientific, industrial, and regulatory needs.

The step 3 of the FAIRification workflow is the most important one, namely definition of a semantic data model for chemical objects representation. In this sense, the efforts for substance data model elucidation are also efforts for FAIR data. The other pillars of a primary importance for the data FAIRness are inclusion of rich metadata (step 6), ontology annotations, and data linking with globally unique identifiers (step 4).

The FAIR principles are combined with so-called CARE (Collective benefit, Power to control, Responsibility, Ethics) principles [22]. CARE principles promote Indigenous Data Management to enhance machine functionality addressing concerns about rights and interests of the Indigenous people in their data throughout the data lifecycle (as a collective to have a say in how their data is actually used), trust, and accountability in the contexts of traditional knowledge and scientific data-oriented toward improving human well-being.

TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) is yet another set of principles [23] aligned with FAIR. To make data FAIR while

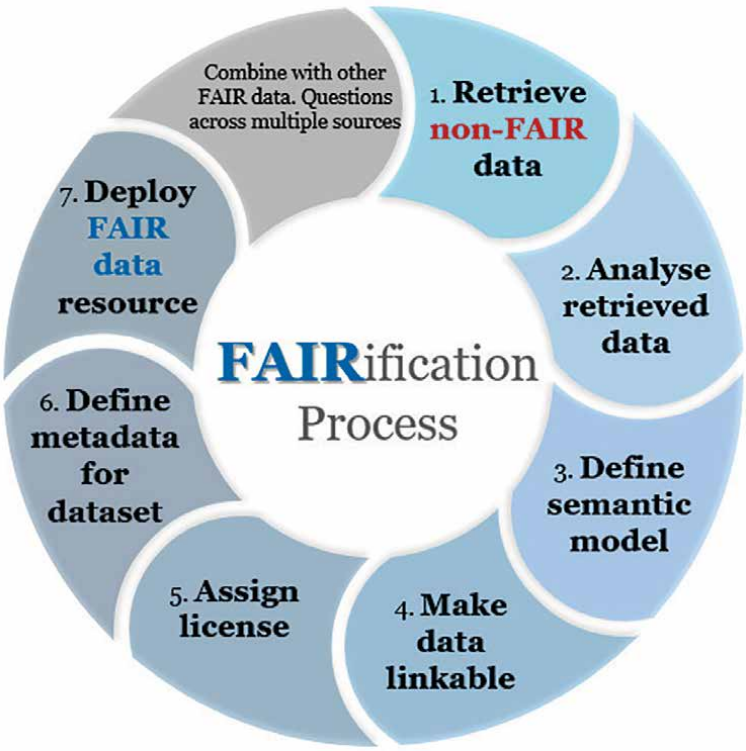


Figure 7.
FAIRification process: a workflow of seven steps for transforming a non-FAIR data into a FAIR data resource.

preserving them over time requires trustworthy digital repositories (TDRs) with sustainable governance and organizational frameworks, reliable infrastructure, and comprehensive policies supporting community-agreed practices. TDRs may actively preserve data within dynamics of technology and stakeholder requirements.
The TRUST principles facilitate communication with all stakeholders, providing repositories and guidance to good practices.

5. Ambit/eNanoMapper data model

Safe by Design approaches are encouraged and promoted through regulatory initiatives and numerous scientific projects. Experimental FAIR data are the basis of risk assessment processing workflows. The Ambit/eNanoMapper [24, 25] database is an open-source chemical data management solution that currently holds the largest compilation of searchable nano-EHS (Environment, Health, and Safety) data in Europe from multiple completed and most of the ongoing H2020 Nano-EHS projects. Ambit is an open-source cheminformatics platform with over 30 modules implemented on the top of CDK [26, 27]. It is funded by CEFIC-LRI (http://cefic-lri.org/) for linking Ambit [28] system with the IUCLID substance database to support read across of substance data, category formation, REST APIs, web interface, substance and structure search facilities, toxicity prediction, and QSAR models. The eNanoMapper database is an extension of the Ambit cheminformatics platform.

The implementation of substance support in Ambit was inspired by the four data models discussed in previous section. The data model has been developed, tested, and improved for about 15 years, processing use cases and feedback from multiple users. The Ambit/eNanoMapper data schema is visualized in **Figure 8**. It contains a variety of data components (entities) serving different roles for the representation of items of information about substances and measurements. The data model entities may have different implementations at different stages of the data processing workflow:

- Serialization on input and output to the system (JSON, RDF or HDF5);
- Java classes in the server side of Ambit implementation;
- Relational database tables at the system back end;
- Python, R, Java, or JavaScript data structures within client libraries.

The data model is a conceptual representation of chemical substances and can be applied with different technologies, enabling interoperability and data linking, internally and externally via REST APIs.

The substances are characterized by their compositions and are identified by names and IDs. The model supports multiple compositions, with one or more components, each with a role assigned. Also, each component is treated via the standard triad approach. The results from physicochemical and biological measurements are treated as properties of the entire substance and are handled via the protocol applications. Efficient experimental protocol description is crucial for the correct communication of the scientific results and for creation of FAIR data resources. The latter is performed by means of a rich set of metadata parameters with a flexible logical organization (e.g. the full experimental data graph defined in ISA data model). The

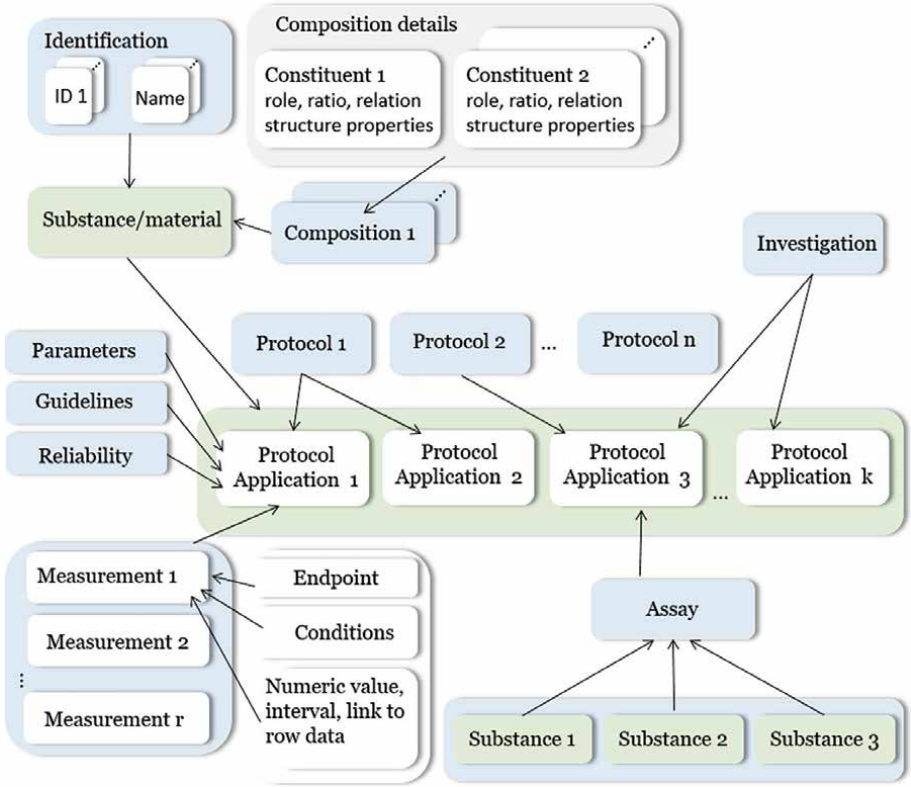


Figure 8.
Schema of the ambit/eNanoMapper data model.

event of applying a test or experimental protocol to a chemical substance is described by a "protocol application" entity. Each protocol application consists of a set of "measurements" for a defined "endpoint" under given "conditions." A measurement result can be a numeric, a string value, or a link to a raw data file (e.g. an IR spectrum, a microscopy image, HTS data, etc.). Measurement entity is also a dynamic data structure where a single number or an interval with lower and upper values together with specified qualifiers are supported. Ambit/eNanoMapper treats miscellaneous cases for a single datum storage, exemplified for the boiling point (BP) endpoint:

BP = 135°C, BP > 130°C, 120°C < BP ≤ 130°C, BP ∼ 3°C, BP = 3 ± 0.5°C.

The measurement errors are represented via a separate qualifier, and different approaches for uncertainty are supported (e.g. SD – standard deviation, MAE – mean average error). The same flexibility is applied for storing metadata parameters. Each measurement is packed with a dynamic list of experimental conditions (or experiment factors such as concentration, time, etc.) which are considered as "lower" level metadata parameters. The "high" level metadata, namely the "protocol application", is described by another dynamic list of parameters, links to Standard Operating Procedures (SOP), guidelines, publications, data quality, etc. The data for a particular substance may contain many "protocol applications."

Figure 9 illustrates different levels of metadata: protocol parameters (Cell type = A549, Method = COMET, and Technical replicates = 3) and varied experimental conditions (Concentration and Exposure time). The same protocol "COMET" can be applied with different parameters (e.g. different cell line and replicates), and

Substance	Protocol application data: Comet	Endpoint	Result	Concentration	Treatment	Exposure time
Substance name: BASO4 NM-220 **Substance UUID:** NRG2-2b94afbf-df44-3f76-ba5b-8973badd91b7 **Public name:** NM-220 **Project:** NanoReg2	**Cell type:** A549 **Exposure time:** 24 h, 3h **Method:** COMET **SOP reference:** link **Input file:** NR2_Scoring data_NILU_20210503.xlsx **Number of technical replicates per conditions:** 3	NET FPG SITES	3.65 % tail	0 ug/ml	sample	24 h
			12.35 % tail	0.16 ug/ml		
			11.59 % tail	0.48 ug/ml		
			12.30 % tail	1.6 ug/ml		
			13.21 % tail	4.8 ug/ml		
			11.99 % tail	16 ug/ml		
			13.75 % tail	48 ug/ml		
			6.94 % tail	120 ug/ml		
			12.31 % tail	160 ug/ml		
			3.29 % tail	0 ug/ml		3 h
			5.82 % tail	0.16 ug/ml		
			5.29 % tail	0.48 ug/ml		
			5.95 % tail	1.6 ug/ml		
			7.32 % tail	4.8 ug/ml		
			6.33 % tail	16 ug/ml		
			15.85 % tail	48 ug/ml		
			12.11 % tail	120 ug/ml		
			13.48 % tail	160 ug/ml		

Figure 9.
Protocol application data: COMET protocol with measurements of endpoint NET FPG SITES, applied for substance NM-220 from public database NanoReg2.

another protocol application will be obtained. The protocol applications that are related to one another are grouped to form an "Investigation" entity. Several different substances that have the same "protocol application" applied can be grouped via the "Assay" entity. The higher level components of the model, such as Substance, Protocol Application, Investigation, and Assay, have automatically generated UUIDs which are used for linking and grouping the measurements.

A transition from the standard triples (S, D, P) to the extended substance data model is challenging for the experts from different domains due to various reasons. Typically the huge volume of metadata compared to the simple experimental data (the ratio of metadata volume to data volume reaches 10:1 or even higher) could be a stumbling block for experimentalists and cheminformatics experts, since a lot of effort is needed for the metadata generation and systematization. However, such a reluctance leads to non-FAIR data and poor findability, accessibility, interoperability, and reusability. Currently, the project funding institutions are challenged in the areas of project results sustainability and reusability as well as the issues of data curation of the results from past research projects and scientific publications.

6. Tools for data FAIRification

Huge volumes of already generated chemical substance data are non-FAIR. One of the predominant ways researchers store their scientific results is in the form of spreadsheets. We demonstrated that the FAIRification can be achieved through the multi-step FAIRification workflow (see **Figure** 7), using the semantic data model of Ambit/eNanoMapper. The analysis of data and metadata is an iterative process, requiring consultations with domain experts to explain the file content and layout, providing SOPs and correct ontology annotations. Generally, the original raw data needs to be converted to the substance data model. For this purpose, a dedicated

software tool was developed, NMDataParser [29], to map the spreadsheets into the Ambit/eNanoMapper semantic model. The latter tool is essentially enabling the most important stage of the FAIRification process – mapping non-FAIR data (e.g. an Excel file) into an existing semantic model.

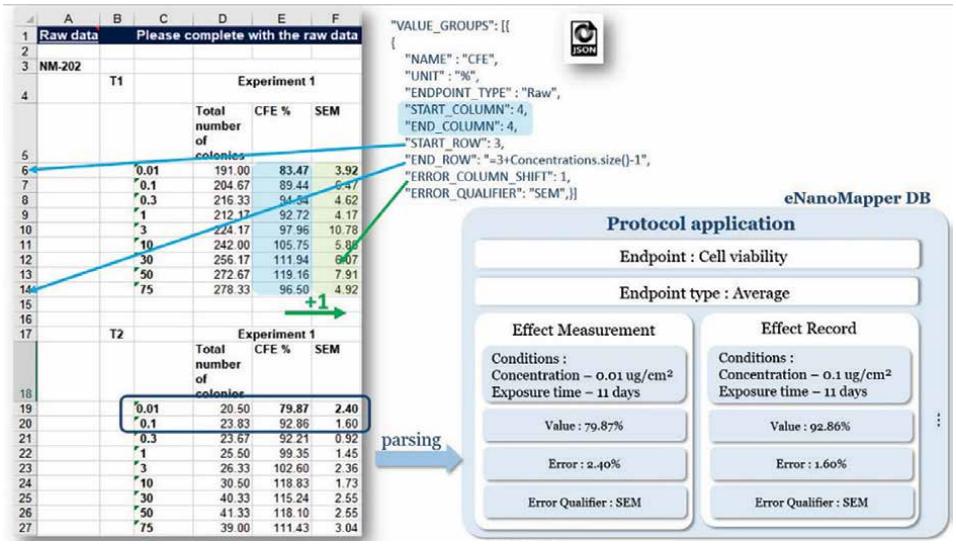


Figure 10.
Parsing of an excel spreadsheet data for a CFE assay of measurements; part of the JSON configuration for relative addressing of the position of the error values is shown (top right); bottom right visualizes the data mapped within the ambit/eNanoMapper substance data model.

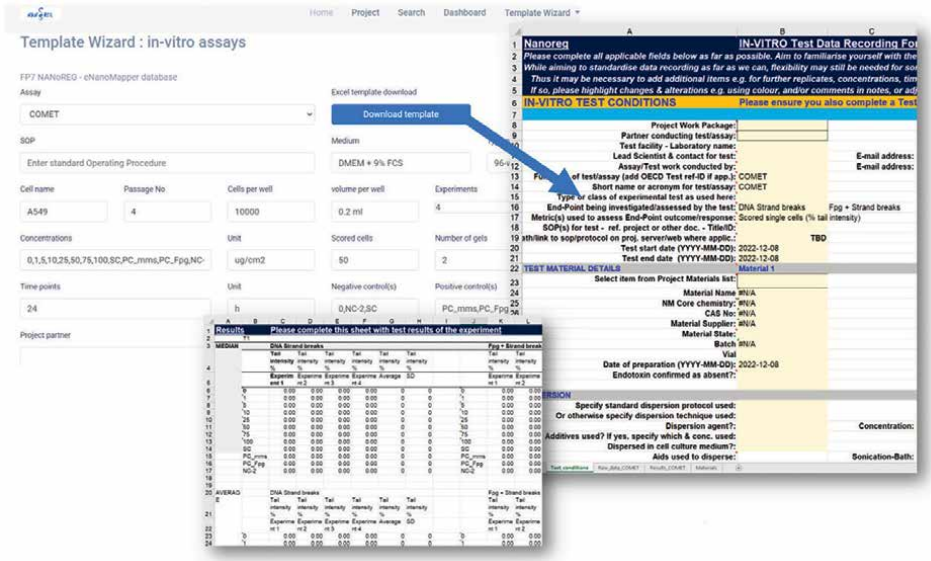


Figure 11.
Web-based template wizard for automatic generation of standardized and harmonized templates with corresponding JSON configuration for NMDataParser.

The NMDataParser is a configurable Excel file parser, developed in Java on top of the Ambit data model and with extensive use of the Apache POI library [30]. It was designed and developed as an open-source library to enable the import of substance data from the Excel spreadsheets with potentially unlimited layout permutations. Different row-based, column-based, or block-based spreadsheet data organizations are supported. The parser is configured via a separate JSON file with its own syntax for mapping the custom spreadsheet structure into the data model components (see **Figure 10**). The parser code, the JSON configuration syntax, documentation, and example files are available at https://github.com/enanomapper/nmdataparser/.

While one JSON configuration file can be applied to multiple Excel files with a similar layout, some complex spreadsheets (e.g. HTS) may require multiple JSON configurations for a single Excel file. The expertise, gained from many years of manual and exhausting configurations of excel file parsing, helped for developing a harmonized and continuously growing set of standard templates which are available via a web interface (see **Figure 11**) with an automatic template generation and corresponding JSON configuration attached.

7. Ambit/eNanoMapper applications, APIs, and services

Once the data is imported into an Ambit/eNanoMapper database instance, it is immediately available (publicly or with a restricted access) via the web user interface and machine readable via an API supporting multiple serialization formats. Non-public datasets are handled by an authentication and authorization system (API keys and OAuth2 plans for direct or delegated access grants are supported). Content from a

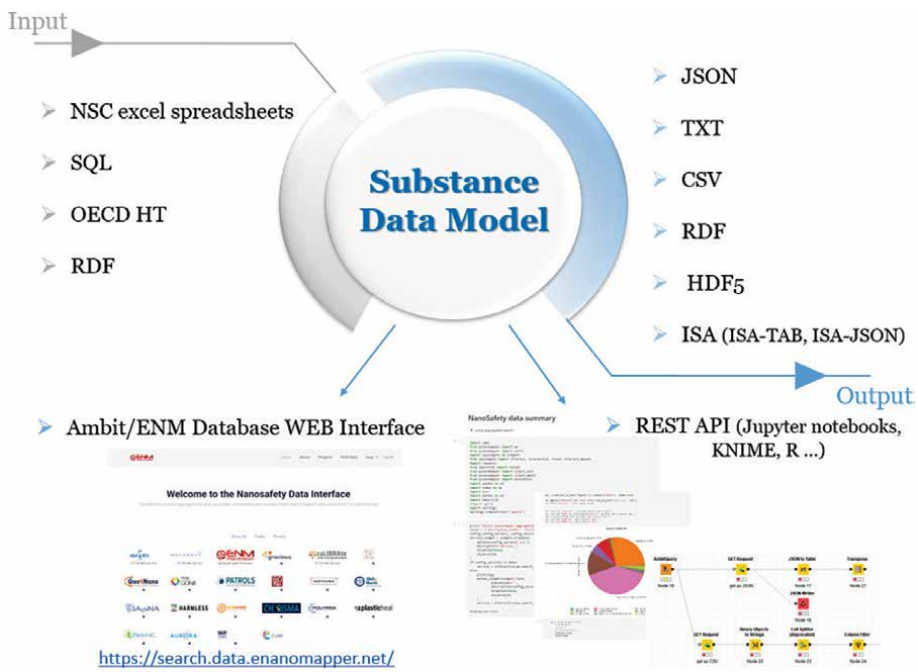


Figure 12.
Data input and output to ambit/eNanoMapper database.

variety of sources such as OECD HTs (IUCLID6 files or direct retrieval from IUCLID servers), custom spreadsheet templates, SQL dumps from other databases, and custom formats, provided by partners (e.g. the NanoWiki RDF dump [31]) is aggregated using the common semantic data model. A variety of options for export, data conversion, data retrieval, and data analysis are available (see **Figure 12**). Different views of substance data are implemented via a Web GUI based on the jToxKit [32] JavaScript library, as well as many customized methods for accessing the data through a REST API via external tools like Jupyter notebooks and the KNIME analytics platform.

Multiple data export formats are supported by the Ambit/eNanoMapper web interface and the API, including semantic formats (RDF and JSON-LD), Excel file formats [28], the native JSON serialization and HDF5 standard.

To facilitate data gap analysis and grouping, the aggregated search interface (see **Figure 13**) includes a number of options, allowing exporting all of the search hits into JSON or spreadsheet formats as well as flexible summary reports in Excel format.

All components of the Ambit/eNanoMapper data model (see **Figure 8**) are searchable. The model schema does not dictate a central entity unlike the standard triad, focused on the chemical structure. This way, more options for searching, storing, and viewing are available. **Figure 13** illustrates a faceted substance search. As it was pointed out, structure searching and substance searching are completely different features of the system. **Figure 14** shows an exact (identity) structure search for the benzene molecule and corresponding logical linking between the molecule of benzene and different chemical substances that contain it as a component.

A number of open-source libraries for accessing the eNanoMapper API are available: https://github.com/enanomapper/renm developed in R; https://github.com/

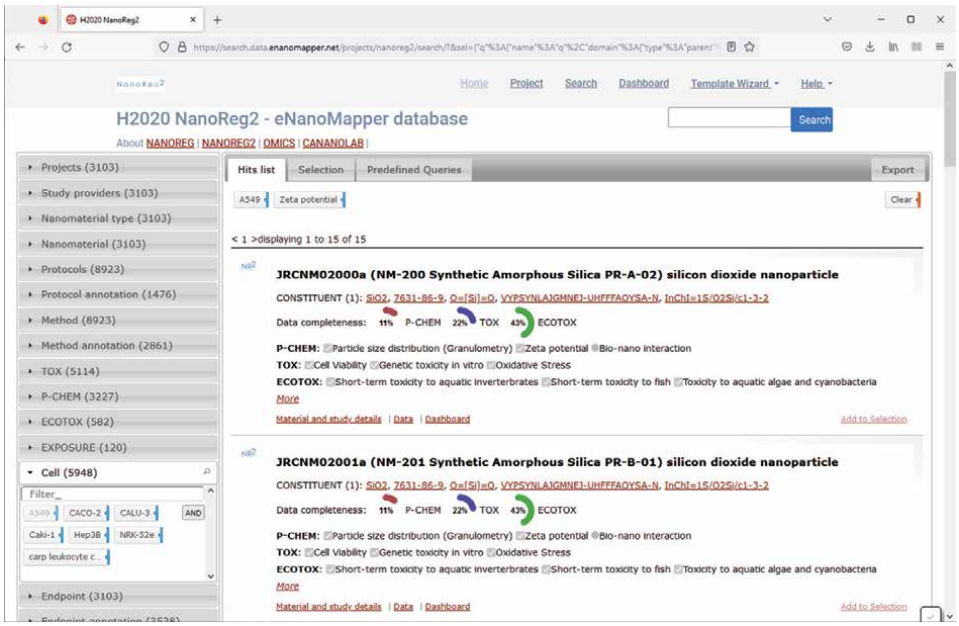


Figure 13.
Faceted search for substances within NanoReg2 public database: Searching for NMs that have experiments with A549 cells and phys-chem characterization with zeta potential.

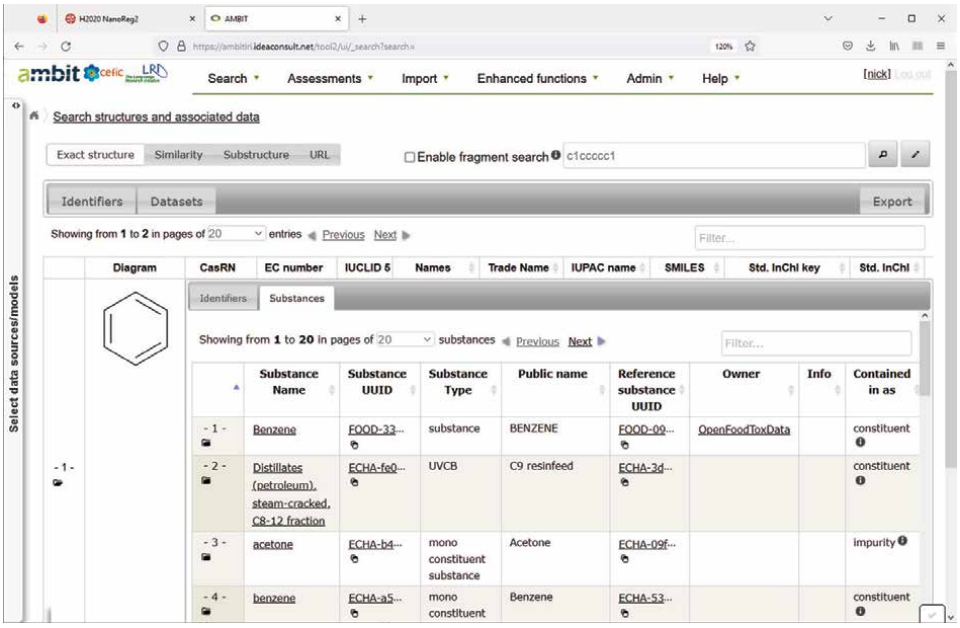


Figure 14.
Exact (identity) structure search in ambit/eNanoMapper database for the benzene structure; result structure is linked to a set of substances having the benzene mole as a component with different roles (rightmost column).

enanomapper/ambit.js and https://github.com/ideaconsult/jToxKit developed in JavaScript; https://github.com/ideaconsult/pynanomapper developed in Python. Python library, in particular, is used for the set of open-source Jupyter notebooks that demonstrate the eNanoMapper API (https://github.com/ideaconsult/notebooks-ambit/tree/master/enanomapper).

8. Linear notations and identifiers for chemical substances

Linear notations are representing chemical structure connectivity and other molecule features as a character string. Linear notations proved to be popular and efficient tools in the field of cheminformatics. The present-day mainstream notations, SMILES [33, 34] and InChI [35], are de facto standards and used in the majority of cheminformatics tools and structural databases. Naturally, linear notations played a significant role for establishing the classical triad model (S, D, P). Linear notation, InChI (International Chemical Identifier), as its name points it out, is originally designed to be a unique structure identifier. The methods for canonical atom numbering and canonical structural presentations are well known (e.g. canonical CTs and canonical SMILES) and together with hashing approaches (e.g. InChI-Key) are widely utilized for structure identification. Also, database and registry molecular numbers are another efficient means for molecule identification. The identification of the chemical substances is a huge challenge, especially in the field of nanoinformatics. Regulatory frameworks experience a lack of unique identifiers, since the traditional identifiers and the most popular linear notations are inadequate. One of the pillars for establishing the FAIR principles is utilization of globally unique and persistent identifiers (see points F1 and F3 from the FAIR principles [3]).

The chemical substance paradigm has been gradually adopted within the cheminformatics and nanoinformatics domains. The substances are serialized via data models with hierarchical organization (e.g. Ambit JSON or ISA model). With a proper canonicalization method, such data serialization (or parts of the data) could be hashed and used as a locally defined identifier, as it is the case with Ambit/eNanoMapper UUIDs (see **Figures 4** and **5**). The complexity of the substance data model justifies the utilization of nonlinear techniques for serialization. However, lately, great effort has been put for developing a linear notation and universal identifiers for chemical substances and NMs.

The InChI Trust (https://www.inchi-trust.org/) works on developing and promoting the use of the IUPAC InChI [35] open-source chemical structure representation algorithm. InChI Trust projects cover versatile types of chemical objects and perform a pioneering work for developing lineation notations for mixtures (MInChI project), nanomaterials (NInChI – project), and Polymer InChI (PInChI – project) – to name a few of the most relevant projects to the chemical substances. Nano-InChI (NInChI) [36] project is a promising effort to integrate concepts of NMs intrinsic and extrinsic properties and to support a domain-specific language for nanoinformatics. NInChI is not intended to replace the chemical substance model but proposes to encode information (composition, size, shape, surface chemistry, etc.) required to unambiguously identify a specific NM as an extension of the IUPAC International Chemical Identifier, termed NInChI. NMs are particulates with specific relationships between the core and surface components that challenges traditional material naming and scientific data communication between researchers, modelers, industry, and regulators. Leveraging best practices with other InChI working groups, e.g. MInChI, Reaction InChI, and PInChI, is planned. NInChI development is a collaborative effort of domain experts from different fields. Currently, the NInChI is under active development, and there are some preliminary NInChI prototypes. For example: Fe3O4 core magnetic nanomaterial with diameter = 38 nm, coated with Glycine and shell thickness of 2 nm can be encodes as:

NInChI = 0.00.1A/C2H5NH2/C3–3-2(4)5/h1,3H2,(H,4,5)/msh/s2t-9!/3Fe.4O/msp/s38d-9/y2&1.

Another possible approach is utilization of SYBYL Line Notation (SLN) [37, 38]. SLN is unambiguous, nonunique linear notation developed by TRIPOS Inc. SLN supports syntax for specification of molecules, substructure queries, and reactions which cover the capabilities of SMILES [33], SMARTS [39], and SMIRKS [40] taken together. On top of the basic syntax, SLN includes other powerful features for the specification of user-defined attributes, macro and Markush [41] atoms for flexible definition of molecular fragments, search queries and structural libraries, as well as 2D and 3D coordinates. All that is accomplished through a unified syntax within a single notation. These features make SLN suitable for data storage and exchange. To our knowledge, SLN is the most comprehensive and rich linear notation for the representation of chemical objects of various kinds facilitating a wide range of cheminformatics algorithms. Though it is not the most popular linear notation nowadays, SLN has excellent capabilities for supporting the challenging tasks of present-day cheminformatics. SLN's rich syntax allows encoding of a comprehensive and versatile chemical information within the boundaries of a linear string representation otherwise manageable with complex data structures such as JSON [21] or XML [42] schemas.

Particularly, SLN is suitable for treating chemical objects with rich metadata (e.g. chemical substances). The SLN string defines one or more fully connected CTs plus a section with molecule attributes for each CT. One of the SLN advantages is the syntax extension, including comparison operations such as <, <=, >, and > =, while the

SMILES/SMARTS standards support only attribute equality. The latter is in line with the substance model flexibility for storing experimental values. Within the existing notations from the past, SLN seems to have the most wide and flexible syntax features to support the chemical substance paradigm. A SLN example for the mentioned above Fe3O4 core magnetic NM, coated with Glycine:

O[1]Fe[2]OFeOFe@1O@2 < role = core;size = 38 nm > CH2(C(=O)OH)NH2 < role = coating;size = 2 nm > .

9. Conclusions

The FAIR principles align with the global shift to open data by promoting governance criteria for increased data sharing. Cheminformatics, nanoinformatics, and bioinformatics methods are providing data-driven solutions in the field of chemical substance safety. The FAIR compliance calls for extension of the structure-centered data models to meet the challenges of chemical substance and materials data management. The substance must include not just a single structure, but a composition of many components with definite roles, corresponding interconnections, rich metadata, and ontology annotations. The variety of data sources, formats, and logical organizations challenges the aggregation of data from multiple projects into a common information system. Ambit/eNanoMapper data model has a well-defined semantics and full adoption of the FAIR principles in order to boost successful strategies for reusable and sustainable research results with efficient interconnections and collaboration between academia, industry, and regulators.

Acknowledgements

The work leading to this chapter has received funding from the European Union's Horizon 2020 Research and Innovation program, Grant Agreements no. 814426 NanoinformaTIX and LRI-EEM9.5 – IC AMBIT.

Author details

Nina Jeliazkova[1*], Nikolay Kochev[1,2] and Gergana Tancheva[2]

1 Ideaconsult Ltd., Sofia, Bulgaria

2 Faculty of Chemistry, University of Plovdiv, Department of Analytical Chemistry and Computer Chemistry, Plovdiv, Bulgaria

*Address all correspondence to: jeliazkova.nina@gmail.com

IntechOpen

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

[1] Gasteger J, Engel T, editors. Chemoinformatics Basic Concepts and Methods. Weinheim: WILEY-VCH Verlag GmbH & Co. KGaA; 2018. p. 575

[2] Massart D, Vandeginste BG, Kaufman L, Demin S, Michotte Y. Chemometrics: A Textbook. Elsevier Science (Verlag); 1988. p. 464. ISBN: 9780080868295

[3] Wilkinson MD, Dumontier M, IjJ A, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Scientific Data. 2016;**3**:1-9. DOI: 10.1038/sdata.2016.18

[4] McNaught AD, Blackwell AW. IUPAC. In: Compendium of Chemical Terminology Chemical Substance. 2014. 2nd ed. Available from: https://goldbook.iupac.org/terms/view/C01039 . p. 2014. DOI: 10.1351/goldbook.C01039

[5] ECHA (REACH). ECHA What is a substance? [Internet]. Available from: https://echa.europa.eu/support/substance-identification/what-is-a-substance. [Accessed: June 12, 2022]

[6] Government of Canada, CEPA. Chemical Substances Glossary [Internet]. 1999. Available from: https://www.canada.ca/en/health-canada/services/chemical-substances/chemical-substances-glossary.html. [Accessed: June 12, 2022]

[7] Epa A. TSCA Chemical Substance Inventory [Internet]. Available from: https://www.epa.gov/tsca-inventory [Accessed: June 12, 2022]

[8] Japan CSCL. Japan CSCL – Chemical Substance Control Law [Internet]. Available from: https://chemical.chemlinked.com/chempedia/japan-cscl-chemical-substance-control-law [Accessed: June 12, 2022]

[9] International Organization for Standardization. ISO/TS 80004-1:2015 - Nanotechnologies – Vocabulary – Part 1: Core-terms. ISO; 2015

[10] The European Commission's Science and Knowledge Service [Internet]. Available from: https://joint-research-centre.ec.europa.eu/index_en [Accessed: June 12, 2022]

[11] Chemicals European Agency in Association with the OECD. IUCLID 6 [Internet]. Available from: https://iuclid6.echa.europa.eu/bg/project-iuclid-6

[12] OECD HT [Internet]. Available from: https://www.oecd.org/ehs/templates/ [Accessed: June 12, 2022]

[13] Abeyruwan S, Vempati UD, Küçük-McGinty H, Visser U, Koleti A, Mir A, et al. Evolving BioAssay ontology (BAO): Modularization, integration and applications. Journal of Biomedical Semantics. 2014; **5**(Suppl. 1):1-22. DOI: 10.1186/2041-1480-5-S1-S5

[14] Visser U, Abeyruwan S, Vempati U, Smith RP, Lemmon V, Schürer SC. BioAssay ontology (BAO): A semantic description of bioassays and high-throughput screening results. BMC Bioinformatics. 2011;**12**:257-273. DOI: 10.1186/1471-2105-12-257

[15] Rumble J, Freiman S, Teague C. Towards a uniform description system for materials on the nanoscale. Chemistry International [Internet].

Available from: https://www.degruyter.com/document/doi/10.1515/ci-2015-0402/html. 2015;**37**(4):3-7. DOI: 10.1515/ci-2015-0402

[16] Rumble J, Freiman S, Teague C. Uniform Description System for Materials on the Nanoscale Prepared by the CODATA-VAMAS Working Group On the Description of Nanomaterials. 2016. Available from: https://zenodo.org/record/56720#.Y48ltMtBxD8

[17] Assunta SS, Rocca-serra P, Field D, Maguire E, Taylor C, Hofmann O, et al. Toward interoperable bioscience data. National Public Grade. 2012;**44**(2): 121-126. DOI: 10.1038/ng.1054

[18] Robinson R, Cronin M, Richarz A, Rallo R. An ISA-TAB-Nano based data collection framework to support data-driven modelling of nanotoxicology. Beilstein Journal of Nanotechnology. 2015;**6**:1978-1999. DOI: 10.3762/bjnano.6.202

[19] Thomas DG, Gaheen S, Harper SL, Fritts M, Klaessig F, Hahn-dantona E, et al. ISA-TAB-Nano: A specification for sharing nanomaterial research data in spreadsheet-based format. BMC Biotechnology. 2013;**13**:2-17. DOI: 10.1186/1472-6750-13-2

[20] ISA-JSON format [Internet]. Available from: https://isa-tools.org/format/specification.html [Accessed: June 12, 2022]

[21] ECMA. JSON (ECMA-404 The JSON Data Interchange Syntax). [Internet]. Geneva, Switzerland: ECMA International. Available from: https://www.ecma-international.org/publications-and-standards/standards/ecma-404/ 2017 [Accessed: June 12, 2022]

[22] Carroll SR, Herczog E, Hudson M, Russell K, Stall S. Operationalizing the CARE and FAIR principles for indigenous data futures. Scientific Data [Internet]. 2021;**8**(1):8-13. DOI: 10.1038/s41597-021-00892-0

[23] Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, et al. The TRUST principles for digital repositories. Scientific Data. 2020;**7**(1):1-5. DOI: 10.1038/s41597-020-0486-7

[24] Jeliazkova N, Apostolova MD, Andreoli C, Barone F, Barrick A, Battistelli C, et al. Towards FAIR nanosafety data. Nature Nanotechnology. 2021;**16**(6):644-654. DOI: 10.1038/s41565-021-00911-6

[25] Jeliazkova N, Chomenidis C, Doganis P, Fadeel B, Grafström R, Hardy B, et al. The eNanoMapper database for nanomaterial safety information. Beilstein Journal of Nanotechnology. 2015;**6**:1609-1634. DOI: 10.3762/bjnano.6.165

[26] Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, et al. The chemistry development kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching. Journal of Cheminformatics. 2017;**9**(1):1-19. DOI: 10.1186/s13321-017-0220-4

[27] Chemistry Development Kit [Internet]. Available from: https://cdk.github.io/ [Accessed: June 12, 2022]

[28] Jeliazkova N, Koch V, Li Q, Jensch U, Reigl JS, Kreiling R, et al. Linking LRI AMBIT chemoinformatic system with the IUCLID substance database to support read-across of substance endpoint data and category formation. Toxicology Letters. 2016;**258**: S114-S115. DOI: 10.1016/j.toxlet.2016.06.1469

[29] Kochev N, Jeliazkova N, Paskaleva V, Tancheva G, Iliev L,

Ritchie P, et al. Your spreadsheets can be fair: A tool and fairification workflow for the enanomapper database. Nanomaterials. 2020;**10**(10):1-23. DOI: 10.3390/nano10101908

[30] Apache POI [Internet]. Available from: https://poi.apache.org/ [Accessed: June 12, 2022]

[31] NanoWiki RDF [Internet]. Available from: https://figshare.com/articles/NanoWiki_4/4141593 2016 [Accessed: June 12, 2022]

[32] JToxKit [Internet]. Available from: https://github.com/ideaconsult/jToxKit [Accessed: June 12, 2022]

[33] SMILES - A Simplified Chemical Language [internet]. Daylight Theory. Available from: https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html [Accessed: June 12, 2022]

[34] Weininger D, Weininger A, Weininger J. SMILES . 2 . Algorithm for generation of unique SMILES notation. Chemical Information and Computer Science. 1989;**29**(19):97-101. DOI: 10.1021/ci00062a008

[35] Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC international chemical identifier [Internet]. Journal of Cheminformatics. 2015;**7**:1-34. DOI: 10.1186/s13321-015-0068-4

[36] Lynch I, Afantitis A, Exner T, Himly M, Lobaskin V, Doganis P, et al. Can an inchi for nano address the need for a simplified representation of complex nanomaterials across experimental and nanoinformatics studies? Nanomaterials. 2020;**10**(12): 1-44. DOI: 10.3390/nano10122493

[37] Ash S, Cline MA, Homer RW, Hurst T, Smith GB. SYBYL line notation (SLN): A versatile language for chemical structure representation. Journal of Chemical Information and Computer Sciences. 1997;**37**(1):71-79. DOI: 10.1021/ci960109j

[38] Homer RW, Swanson J, Jilek RJ, Hurst T, Clark RD. SYBYL line notation (SLN): A single notation to represent chemical structures, queries, reactions, and virtual libraries. Journal of Chemical Information and Modeling. 2008;**48**(12): 2294-2307. DOI: 10.1021/ci7004687

[39] SMARTS - A Language for Describing Molecular Patterns [Internet]. Daylight Theory. Available from: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html [Accessed: June 12, 2022]

[40] SMIRKS - A Reaction Transform Language [Internet]. Daylight Theory. Available from: https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html [Accessed: June 12, 2022]

[41] Barnard J, Wright PM. Towards in-house searching of Markush structures from patents. World Patent Information. 2009;**31**(2):97-103. DOI: 10.1016/j.wpi.2008.09.012

[42] Extensible Markup Language (XML) 1.0 (Fifth Edition) [Internet]. 2008. Available from: https://www.w3.org/TR/REC-xml/ [Accessed: June 12, 2022]

Section 3

Data Governance and Applications

Chapter 5

Ethical Considerations for Health Research Data Governance

Mantombi Maseme

Abstract

Research involving humans often generates considerable data irrespective of the context in which the research is being conducted. This data must be protected from unauthorized access, use, and sharing as a means of safe-guarding research participants' rights. Notwithstanding the fact that several jurisdictions globally have promulgated laws and regulations aimed at protecting individual citizens' personal information, violation of privacy and related rights occurs in some instances. This could partly relate to a general lack of health research sector specific data governance policies and laws, which include data transfer agreements prevalent in most countries. The chapter therefore aims to cover the ethical aspects of health research data access, use, and sharing as a means of enabling health research institutions and policymakers to develop robust data governance structures and procedures. The scope of the chapter covers health research data generated in empirical research as well as that which is produced within a medical laboratory research context, i.e., human sample associated data.

Keywords: data access, data use, data sharing, data governance, privacy, confidentiality

1. Introduction

Data governance is defined as "all processes related to the collection, storage, processing, curation, use, and deletion of data" [1]. Data governance entails not only the development and rules for data quality management but also specifying the responsibility for making decisions related to data handling as well as the duties related to such decisions [2]. Data governance also assures compliance with the laws governing data [3]. Notably, there is a presumption that data governance is a universal approach with a one size that fits all organizations alike. Weber et al. argue that this should not be the case [3]. Accordingly, data governance within the health research context is considered in this chapter. In the context of this chapter, health research data governance refers to the development of structures and processes for the access, use, and sharing of health research data. The question of why data governance matters in the context of health research is that it is mainly for the purpose of safe-guarding the individual data subject's rights by the data custodians. Infamous cases of unauthorized health research data access, use, and sharing have been well documented [4, 5]. This is despite the existence of regulations aimed at protecting individual citizens' personal information in certain countries. To demonstrate the issue of health research data use

that is not in line with consent granted, health research data misuse is discussed in this chapter. A key consideration for using personal information in medical research is to seek informed consent [6]. Accordingly, health research data and consent will be discussed in this chapter.

2. Governance of health research data

According to the WMA Declaration of Taipei, governance of health databases should be based on the principles of [7]: (1) protection of individual rights' over the interests of other stakeholders of science; (2) transparency in making any relevant information available to the public; (3) participation and inclusion of individuals and communities by health database custodians through consultations and engagement as well as (4) accountability in that custodians of health databases should be accessible to all stakeholders. Correspondingly, each of these principles for data governance is discussed in turn in subsequent sections.

The individual (human participant) rights to data (interchangeably health research data herein) access that should be respected include the right to privacy and confidentiality, notably, anonymization or confidentiality measures for ensuring confidentiality should be considered by the data custodians [8]. Other applicable rights that ought to be observed with respect to data access, use, and sharing include the right to autonomy and the right to dignity [7]. Human rights should be protected by the rule of law [9] and as such human rights are inalienable (should not be taken away) except in specific circumstances, e.g., restriction of liberty if a person is found guilty of committing a crime by a court of law [10]. The CIOMS ethical guidelines also recommend governance systems that uphold the principle of accountability while maintaining good stewardship for samples and their associated data [11].

Some of the elements for governance of health data (including health research data) include, *inter alia*: arrangements for duration of data storage; nature and purpose of data collection; arrangements for disposal and destruction of data; arrangement for dealing with the data in the event of change of ownership, obtaining consent, protecting the rights of data subjects (interchangeably research participants herein); criteria and procedures for data access and sharing as well as measures to prevent unauthorized access or inappropriate sharing as well as; responsibility for data governance [7].

2.1 Protection of data subjects' rights

As already mentioned, the rights that must be protected in relation to access, use, and sharing of personal information are the rights to: privacy and confidentiality; autonomy and dignity.

Confidential information is that which is sensitive, needs protection, and can only be disclosed in a trusting relationship [12]. A recommendation for maintaining confidentiality of health research data is through anonymization or coding in order to protect the participant (data subject) from harm or stigmatization [11]. Moreover, confidentiality is a significant standard of professionalism [12]. Privacy on the other hand is difficult to define, with no universally accepted definition and has been subject to extensive debate by philosophers and legal scholars [13]. In modern society, however, the term addresses the question of who has access to personal information and the conditions of such access [13]. Primary justification for protecting privacy of

persons is to protect their interests [13]. Based on this notion, it follows then that ethical justification for protecting privacy of persons aligns with the notion of respecting participant's autonomy (self-determination) by virtue of the objective of privacy being to protect individual interests.

Professional codes of practice for healthcare professionals sometimes appeal to norms of respecting autonomy and privacy as well as not harming others (non-maleficence) although these norms are not explicitly expressed in such codes [14]. The rights to autonomy, privacy, and confidentiality call for not only protection of bodily integrity, but also that the scope of decision-making should be free from interference by others [15]. It can therefore be inferred that access and use of health research data that are outside the scope of the data subject's decision-making for participation (consent) violate these rights. Some of the factors that promote noncompliance of the principles of ethics in health research include lack of ethical supervision; paternalism caused by trust in the researcher, resulting in a subtle loss of participant autonomy; informed consent documents that do not adequately address all aspects of research participation, in particular the potential risks involved, thereby threatening autonomy and engendering maleficence if there are any risks involved; lack of legislation to protect participants and pressure for researchers to increase research output at the expense of ethical principles [16].

2.2 Transparency in data governance

Smits and Champagne define transparency as the disclosure of procedure and results [17]. Mahsa and Mojisola consider transparency to be "the development and public availability of data sharing and access policies" [18]. Notably, there is no single conception of transparency, but rather it has multiple definitions, purposes, and applications with users of transparency including self-governing citizens, governments, and private firms [19]. Transparent governance for health data requires clear information for data circulation, data-sharing agreements, research objectives, and findings to be made available to the public [20]. In the context of biobank governance, transparency enables donors to better understand biobank governance and thereby make better, more informed decisions about sample and associated data donation as well as research participation [21]. Lack of transparency has the potential to undermine public trust in initiatives that involve large datasets such as biobanks [21].

A notable limitation of making information available to the public is that of web based transparency particularly for communities that lack web access or information technology infrastructure [21]. If data collection or use raises specific ethical questions, e.g., with regard to consent and transparency as well as privacy and data subjects' rights and expectations, an explanation of how the ethical concerns will be mitigated is requisite in the operational plan for data collection and processing [22]. Community engagement has been identified as an important element of ethical research data sharing by research stakeholders that include researchers and health providers, community representatives, assistant chiefs, and field workers [23].

2.3 Community and individual engagement in the context of health research data governance

Community (public) engagement is perceived as a means of cultivating public trust and cooperation in research activities [24]. Ethical justification for community

engagement is that it improves the consent process, identifies ethical issues and develops processes for resolving ethical issues when they arise [25]. Jao et al. identified a number of community engagement goals in relation to health research data access, and these include: (1) creating an awareness of data sharing activities with information on how any benefits and risks would be managed; (2) giving feedback to the community or representatives on the data sharing process; (3) ad hoc community consultation in relation to specific data sharing requests [23]. Evans et al. propose Community Advisory Boards (CABs) as platforms to engage the affected populations on how they would want their data to be collected, stored, and shared [26]. The Mayo Clinic Biobank CABs provide a way of incorporating interested community groups in the governance of large-scale bio-resources [27].

Engagement of research participant groups as well is important because: participants are in a better position to speak to the risks associated with the use of their data as well as having an interest in the use of their data, which the public might not necessarily have [27]. Moreover, engaging participants in data management decisions has been cited as strengthening transparency and accountability [28].

2.4 Accountability for health research data

The WMA defines accountability as that which "requires being prepared to provide an explanation for something one has done or has not done" [7]. Accountability represents a moral obligation to answer and the practical ability to convey that answer [29]. Moreover, accountability serves to establish responsibilities [29]. In the context of health research data governance, it is the responsibilities of the data custodians that should be established. This chapter refers to data custodianship rather than ownership because firstly, claiming ownership rights of data is a misconception in that proprietary rights over data do not exist in the international intellectual property (IP) system [30]. Secondly, because trends on claims of data ownership are based on "flawed models and on implausible arguments" [30]. Data custodians are also referred to as stewards [31]. Accordingly, data steward(ship) and custodian(ship) will be used interchangeably in this chapter. The concept of a data steward is intended to denote a level of fiduciary (trust) responsibility toward the data [32]. Moreover, responsibilities for data stewardship are conceptualized and fulfilled by the process of governance [32]. Data stewardship entails the existence of mechanisms for the responsible acquisition, storage, safe-guarding, and use of data [32]. The concept of custodians should also ensure the existence of systems that ensure privacy of individuals at every stage [33] as well as ensuring an adequate level of confidentiality of such data in order to preserve the data as much as possible for the researchers [8].

Empirical research conducted in Australian data custodians shows that they perceive their role to be more of protecting data subjects' privacy than other vulnerabilities [31]. Data custodians also have the responsibility of ensuring that data sharing complies with legal and policy requirements prior authorizing the release of data on behalf of institutions [31].

3. Ethical considerations and issues in health research data access, sharing, and use

There is a wide recognition that sharing of data generated from research involving humans raises ethical and governance issues [34]. Some of the issues (risks) of

data access and sharing include: (1) confidentiality and privacy breaches as well as the need to manage these two aspects and (2) violation of expectations of data reuse [35]. These are ethical challenges because of their potential to violate human dignity and autonomy as well as pose a risk of discrimination [35]. Other ethical considerations for data sharing include: valid consent particularly when future uses of data are unclear; the potential impact of such sharing on public trust and implication for future research in terms of inappropriate data use, e.g., publication of data in discriminatory ways and issues related to decisions on data access [36]. According to interviewees in a study on health research data ethical practices conducted in South Africa (SA), researchers have an ethical duty to provide accurate data as a means of nurturing professional integrity through transparent practice, coupled with avoidance of unauthorized future research use [37]. Such stakeholder views should be addressed by policies to ensure ethical data sharing nationally and internationally [37]. The noted ethical issues pertaining to health research data access, sharing, and use are discussed in turn in the subsequent sections.

3.1 Health research data breaches

Sharing of research data requires adequate safeguards for the protection of participants' rights and should also be fully consistent with the terms of consent granted [38]. Unauthorized users are able to access databases due to vulnerabilities in software, human error, and security failures resulting in sensitive data being exposed leading to confidentiality breaches [39]. Lord et al. argue that using anonymized health research data need not be regarded as a confidentiality breach claiming that informed consent is unnecessary and often impractical [40]. The basis for this claim is unclear but seems to be based on the notion that anonymized data use benefits outweigh any confidentiality issues. Anonymization is the process of irreversibly removing from a dataset those variables that can identify an individual [41]. In the context of this chapter, health research data misuse refers to access, sharing, and use of such data that is not in line with consent granted. There is paucity of literature on misuse of health research data; however, a case in point is that of an alleged report of African sample associated data that was transferred from SA to the UK to develop gene chips [5]. Neither SA researchers nor research participants were aware of such purported commercialization of African health research data [5]. If proven to be true, such allegations demonstrate a violation of research participants' autonomy and dignity. When data are not anonymized, the risk of malevolent exploitation seems to be significantly increased [42].

Iceland Health Sector Database (HSD) legislation and the visibility of its processes have exposed the innovation of genomics to a public debate resulting in exposure of ethical issues of commodification of bioinformatics (the fusion of biotechnology and informatics) and human tissue to the international cultural and political agenda [43]. Nicolson argues that data reuse by healthcare professionals and researchers commodifies people's medical records and reduces such data to a commodity that can be bought and sold based on the reasoning that data reuse may reinforce social inequalities [44]. This argument does not hold, particularly when data reuse is in line with ethico-legal requirements.

Public forum comprising of respondents from professions in legal, ethics, medicine, medical, and social scientists, government professional, security, digital health, and bioinformatics in a study by Staunton et al. in 2019 in SA had a general awareness of the need for protection of personal health information [45].

3.2 Health research data reuse and consent

It is ethically mandatory for the data subject's rights to be protected [46]. Meystre et al. propose principles for ethical data reuse, and these include principles of: information (privacy and disposition—right to privacy and control the use of one's data); openness (appropriate and timely data disclosure); security (data protection through appropriate measures); least intrusive alternative (any violation of privacy or individual's right or control his/her data may occur in a least intrusive manner with minimal interference of the person's rights and accountability (infringement of rights and control of an individual's data must be justified to the affected individual in a timely and appropriate manner) [46]. A significant number of research participants across empirical research studies prefer to be contacted and re-consented for the reuse of their data [47]. The majority of Quebec citizens in a study by Cumyn et al. expressed support for the reuse of health data provided that individuals are informed about such use and consent is sought [48]. Reusing health data without informed consent contravenes patients' expectations resulting in violation of the patients' perceived ownership rights [49]. Moreover, autonomy is a fundamental human right, which may be limited during public health emergencies, provided that such an interference is deemed necessary [49]. Key issues in the discussion about limits for the use of personal data in medical research relate to the scope and limitations of consent as a legal basis for such use [50]. Moreover, one of the principles for processing of personal data in the European Union (EU) regulatory framework is lawfulness, which mandates consent or another legitimate basis (laid down by the law) as requirements for such processing [50].

As already mentioned, another ethical concern is when future uses of data are unclear; accordingly, the potential impact of such sharing on public trust follows in the next section.

3.3 The potential impact of unclear purpose of health data sharing on public trust

There is a long-standing doctor-patient trust relationship through which the doctor (or researcher in this context) is bound by professional integrity to act in the best interests of their patients (or research participants in this context) [51]. Kerasidou submits that trust is important in biomedical research and that professional integrity can promote trust in research [52]. The presence of legally binding ethico-regulatory frameworks aimed at protecting the dignity of research participants enables the development of researcher-participant trust [53]. Trust is an essential element of building and maintaining mutual respect, particularly in relationships where there is an imbalance of power [54]. Kraft et al. have identified factors that influence research participants' trust, and these include: (1) participants' varying benefits expectations, (2) historical discrimination in research, (3) participants' fear that their data might be used inappropriately [55]. These factors will be explored in detail in subsequent sections. The latter factor aligns with health data breaches discussed in Section 3.1 and will therefore not be discussed further in this section. Trust issues in medical research are also caused by exploitation of vulnerable populations, different regulatory frameworks, particularly in research collaborations as well as lack of robust operational management particularly of biobanks as cited by SA researchers in a study conducted by Moodley et al. [56].

Trust relationship between researchers and participants is built when researchers share information, reciprocity based on integrity and equality in replacing

vulnerability and dependence [57]. It is not only prospective research information that can be shared with participants but also research findings through community (public) engagement as a means of building trust [58]. Public engagement has been identified as a key mechanism for building trust [59]. Another requirement for building the healthcare professional-patient trust relationship is confidentiality [60]. Therefore, trust imposes a duty of maintaining confidentiality on healthcare professionals.

3.3.1 Research participants' varying benefits expectations and trust

Molyneux et al. distinguish between direct benefits for research participants (e.g., diagnostic tests, distribution of medication, evaluation services) and indirect (collateral) benefits for those that are not specifically targeted at research participants but might include research participants as well (e.g., provision of healthcare services to family or community members) in a fair benefits framework [61]. A study conducted in Kenyan views and experiences on research participants' benefits and payments showed that inconsistencies in research benefits such as varying transport fares on different occasions for the same study has the potential for introducing participant mistrust [61]. Biobank research participants in a qualitative study in Australia declared institutional trust in that they did not necessarily trust the individual researchers but rather the research institution based on a perception that the institution carrying out the research was reputable [62]. Some researchers, particularly in the social and behavioral sciences, believe that research participant deception that does not involve any harm is justified, while those who oppose this view hold that such deception violates and takes advantage of participants' trust in scientists [63].

Benefit sharing in research, particularly in genetic data banking, is recommended not only from an ethical obligation point of view but also as a potential solution to resolving the issue of loss of public trust in a sense that benefit sharing is recognition for participants' contribution to the research endeavor [64]. Nicol and Critchley note that based on the notion of reprocity, biobanks, which reward contribution, through benefit sharing, should promote trust, which in turn leads to public participation in biomedical research [65]. Johnson et al., however, are of the opinion that patients are expected to participate in research without benefitting personally because research is a requirement for good-quality healthcare [66].

3.3.2 Historical discrimination in research

Literature on historical discrimination in research involving humans is dominated by the infamous Tuskegee study [67–69]. Briefly, the study that commenced in 1932 in the USA involving a total of 600 black men of which 399 had syphilis and 201 did not have the disease involved a number of ethical transgressions, which were inflicted on the research participants [70]. The ethical violations included participant consent not being sought as well as participants not being adequately informed in that they were misled on the purpose of sample collection under the guise that they were being treated for "bad blood" in exchange for free medical examinations, meals, and burial insurance. Another ethical transgression was the maleficence inflicted on the participants in that even when penicillin became widely available as a treatment of choice in 1943, the participants were not offered treatment [70]. By the time the study was terminated in 1972, after having being leaked by the press, out of the 399 participants who had syphilis, 28 had died, another 100 from syphilis-related complications, 40

patients' wives contracted the disease, and 19 children were infected at birth [71]. Lack of trust regarding the healthcare system and health researchers as cited by research participants, particularly African Americans, has historical roots, with the Tuskegee syphilis study having been cited either explicitly or implicitly and its impact continues throughout the generations [67]. Perceived discrimination contributes to higher societal distrust of African Americans in the healthcare system compared with their white counterparts [68]. Such historically nuanced concerns should be addressed by institutional review boards (IRBs) even though the process may be frustrating because such assessments are imprecise by nature [72]. An interesting finding by Brandon et al. revealed that black race, not necessarily knowledge about the Tuskegee study, was a predictor of medical care mistrust, and it is believed that African American mistrust arises from a general mistrust of societal institutions with the Tuskegee study being confirmation of what is speculated or already known about African American treatment in medical systems [69].

The Nuremberg trials involved specific crimes that took place during World War II in which German physicians conducted a series of more than 12 medical experiments in concentration camp inmates with some of the crimes involving killing of Jews for anatomical research, euthanasia of sick and disabled civilians, and killing of tubercular Poles [73].

4. Conclusion

This chapter has considered health research data governance through the lens of the WMA Declaration of Taipei that is based on requirements for: (1) protection of participants' rights; (2) transparency of information; (3) individual and community inclusion through engagement; and (4) accountability of health database custodians through being accessible to all stakeholders. The ethical considerations for health research data access, sharing, and use include: confidentiality and privacy breaches as well as the need to manage these two aspects; violations of expectations of data reuse; valid consent and the potential impact of such sharing on public trust. Factors that influence research participants' trust include: (1) varying benefits expectations, (2) historical discrimination in research, (3) participants' fear that their data might be used inappropriately. These factors have the potential to erode trust of health research data subjects because the context is the same, i.e., research.

Conflict of interest

The author declares no conflict of interest.

Acronyms and abbreviations

HSD	Iceland Health Sector Database
OECD	Organisation for Economic Co-operation and Development
SA	South Africa
WMA	World Medical Association

Author details

Mantombi Maseme
National Health Laboratory Service, Johannesburg, South Africa

*Address all correspondence to: masememr@gmail.com

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

[1] Stahl BC, Rainey S, Harris E, Fothergill BT. The role of ethics in data governance of large neuro-ICT projects. Journal of the American Medical Informatics Association. 2018;**25**(8): 1099-1107. DOI: 10.1093/jamia/ocy040

[2] Otto B. Organizing Data governance: Findings from the telecommunications industry and consequences for large service providers. Communications of the Association for Information Systems. 2011;**29**(3):45-66. Available from: http://aisel.aisnet.org/cais/vol29/iss1/3?utm_source=aisel.aisnet.org%2Fcais%2Fvol29%2Fiss1%2F3&utm_medium=PDF&utm_campaign=PDFCoverPages

[3] Weber K, Otto B, Österle H. One size does not fit all—A contingency approach to data governance. ACM Journal of Data and Information Quality. 2009;**1**(1):1-27. DOI: 10.1145/1515693.1515696

[4] Caulfield T, Burningham S, Joly Y, Master Z, Shabani M, Borry P, et al. A review of the key issues associated with the commercialization of biobanks. Journal of Law and the Biosciences. 2014:94-110. DOI: 10.1093/jlb/lst004

[5] Moodley K, Kleinsmidt A. Allegations of misuse of African DNA in the UK: Will data protection legislation in South Africa be sufficient to prevent a recurrence? Developing World Bioethics. 2020;**00**:1-6. DOI: 10.1111/dewb.12277

[6] Singleton P, Wadsworth M. Confidentiality and consent in medical research: Consent for the use of personal medical data in research. BMJ. 2006;**333**:255-258

[7] World Medical Association. WMA Declaration of Taipei on Ethical Considerations Regarding Health Databases and Biobanks. 2016. Available from: https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/. [Accessed: June 14, 2022]

[8] Organisation for Economic Co-operation and Development. OECD Principles and Guidelines for Access to Research Data from Public Funding. 2007. Available from: https://www.oecd.org/sti/inno/38500813.pdf. [Accessed: June 14, 2022]

[9] United Nations. Universal Declaration of Human Rights. 2015. Available from: https://www.un.org/en/udhrbook/pdf/udhr_booklet_en_web.pdf. [Accessed: June 14, 2022]

[10] United Nations Office of the High Commissioner. What are Human Rights. 2022. Available from: https://www.ohchr.org/en/what-are-human-rights. [Accessed: June 14, 2022]

[11] Council for International Organizations of Medical Sciences. International Ethical Guidelines for Health-related Research Involving Humans. 2016. Available from: https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf. [Accessed: June 14, 2022]

[12] DeJong. Chapter Four-Confidentiality. 2014. Available from: https://www.sciencedirect.com/science/article/pii/B9780124081284000047. [Accessed: June 14, 2022]

[13] Nass SJ, Nass SJ, Levit LA, Gostin LO. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research. Washington

DC: The National Academies Press; 2009. 76 p. Available from: http://www.nap.edu/catalog/12458.html

[14] Beauchamp TL, Childress JF. Moral and ethical theory. In: Beauchamp TL, Childress JF, editors. Principles of Biomedical Ethics. 4th ed. New York: Oxford University Press; 1994. p. 6

[15] Beauchamp TL, Childress JF. Liberal individualism: Rights-based theory. In: Beauchamp TL, Childress JF, editors. Principles of Biomedical Ethics. 4th ed. New York: Oxford University Press; 1994. p. 6

[16] Moreno BAC, Arteaga GMG. Violation of ethical principles in clinical research. Influences and possible solutions for Latin America. BMC Medical Ethics. 2012;**13**(35):1-4. Available from: http://www.biomedcentral.com/1472-6939/13/35

[17] Smits P, Champagne F. Governance of health research funding institutions: An integrated conceptual framework and actionable functions of governance. Health Research Policy and Systems. 2020;**18**(22):1-19. DOI: 10.1186/s12961-020-0525-z

[18] Shabani M, Obasa M. Transparency and objectivity in governance of clinical trials data sharing: Current practices and approaches. Clinical Trials. 2021:45-60. DOI: 10.1177/1740774519865517

[19] Kosack S, Fung A. Does transparency improve governance? Annual Review of Political Science. 2014;**17**:65-87

[20] Ford E, Boyd A, Bowles JKF, Havard A, Aldridge RW, Curcin V, et al. Our data, our society, our health: A vision for inclusive and transparent health data science in the United Kingdom and beyond. Learning Health Systems. 2019;**3**(e10191):1-12. DOI: 10.1002/lrh2.10191

[21] Gille F, Axler R, Blasimme A. Transparency about governance contributes to biobanks' trustworthiness: Call for action. Biopreservation and Biobanking. 2020:1-2. DOI: 10.1089/bio.2020.0057

[22] European Commission. Ethics and Data Protection. 2018. Available from: https://ec.europa.eu/info/sites/default/files/5._h2020_ethics_and_data_protection_0.pdf. [Accessed: June 14, 2022]

[23] Jao I, Kombe F, Mwalukore S, Bull S, Parker M, Kamuya D, et al. Involving research stakeholders in developing policy on sharing public health research data in Kenya: Views on fair process for informed consent, access oversight, and community engagement. Journal of Empirical Research on Human Research Ethics. 2015;**10**(3):264-277. DOI: 10.1177/1556264615592385

[24] Sleigh J, Vayena E. Public engagement with health data governance: The role of visuality. Humanities and Social Sciences Communications. 2021;**8**(149):1-12. DOI: 10.1057/s41599-021-00826-6

[25] National Institutes of Health. Community Engagement. 2011. Available from: https://www.atsdr.cdc.gov/communityengagement/pdf/PCE_Report_508_FINAL.pdf. [Accessed: June 18, 2022]

[26] Evans EA, Delorme E, Cyr K, Goldstein DM. A qualitative study of big data and the opioid epidemic: Recommendations for data governance. BMC Medical Ethics. 2020;**21**(101):1-13. DOI: 10.1186/s12910-020-00544-9

[27] Milne R, Sorbie A, Dixon-Woods M. What can data trusts for health research learn from participatory governance in biobanks? Journal of Medical

Ethics. 2022;**48**:323-328. DOI: 10.1136/medethics-2020-107020

[28] Shah N, Coathup V, Teare H, Forgie I, Giordano GN, Hansen TH, et al. Sharing data for future research—Engaging participants' views about data governance beyond the original project: A DIRECT study. Genetics in Medicine. 2019;**21**(5):1131-1138

[29] Hoeyer K, Bauer S, Pickersgill M. Datafication and accountability in public health: Introduction to a special issue. Social Studies of Science. 2019;**49**(4):459-475. DOI: 10.1177/0306312719860202

[30] Andanda P. Towards a Paradigm Shift in Governing Data Access and Related Intellectual Property Rights in Big Data and Health-Related Research. Springer. 2019;**50**:1052-1081. DOI: 10.1007/s40319-019-00873-2

[31] Allen J, Adams C, Flack F. The role of data custodians in establishing and maintaining social licence for health research. Bioethics. 2019;**41**:404-409. DOI: 10.1111/bioe.12549

[32] Rosenbaum S. Data governance and stewardship: Designing data stewardship entities and advancing data access. Health Services Research. 2020;(Special Issue):1442-1455. DOI: 10.1111/j.1475-6773.2010.01140.x

[33] United Nations. 2018. A Human Rights-based Approach to Data. Available from: https://www.ohchr.org/sites/default/files/Documents/Issues/HRIndicators/GuidanceNoteonApproachtoData.pdf. [Accessed: June 24, 2022]

[34] Wellcome Trust. 2015. Ethical Sharing of Health Research Data in Low- and Middle-income Countries: Views of Research Stakeholders. Available from: https://cms.wellcome.org/sites/default/files/ethical-sharing-of-health-research-data-in-low-and-middle-income-countries-phrdf-2014.pdf [Accessed: June 23, 2022]

[35] Organisation for Economic Co-Operation and Development. OECD iLibrary. 2019. Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies. Available from: https://www.oecd-ilibrary.org/sites/15c62f9c-en/index.html?itemId=/content/component/15c62f9c-en. [Accessed: June 23, 2022]

[36] O'Connell and Plewes. 2015. Sharing Research Data to Improve Public Health in Africa. Available from: https://www.ncbi.nlm.nih.gov/books/NBK321547/pdf/Bookshelf_NBK321547.pdf. [Accessed: June 23, 2022]

[37] Denny SG, Silaigwana B, Wassenaar D, Bull S, Parker M. Developing ethical practices for public health research data sharing in South Africa: The views and experiences from a diverse sample of research stakeholders. Journal of Empirical Research on Human Research Ethics. 2015;**10**(3):290-301. DOI: 10.1177/1556264615592386

[38] Sankoh O. Sharing research data to improve public health. The Lancet. 2011;**377**:537-539. DOI: 10.1016/S0140- 6736(10)62234-9

[39] Seh AH, Zarour M, Alenezi M, Sarkar AK, Agrawal A, Kumar R, et al. Healthcare data breaches: Insights and implications. Healthcare. 2020;**8**(133):1-18. DOI: 10.3390/healthcare8020133

[40] Lord W, Doll R, Asscher W, Hurley R, Langman M, Gillon R, et al. Consequences for research if use of anonymised patient data breaches confidentiality. BMJ. 1999;**319**:1366-1372

[41] World Health Organisation. 2022. Sharing and Reuse of Health-related Data for Research Purposes: WHO Policy and Implementation Guidance. Available from: https://www.who.int/publications/i/item/9789240044968. [Accessed: July 04, 2022]

[42] Ostherr K, Borodina S, Bracken RC, Lotterman C, Storer E, Williams B. Trust and privacy in the context of user-generated health data. Big Data and Society. 2017:1-11. DOI: 10.1177/2053951717704673

[43] Rose H. The Commodification of Bioinformation: The Icelandic Health Sector Database. London: The Wellcome Trust; 2001. p. 31

[44] Nicolson J. The commodification of patient medical records. BMJ. 2013;**347**(f5867):1-1. DOI: 10.1136/bmj.f5867

[45] Staunton C, Tschigg K, Sherman G. Data protection, data management, and data sharing: Stakeholder perspectives on the protection of personal health information in South Africa. PLoS One. 2021;**16**(12):1-19. DOI: 10.1371/journal.pone.0260341

[46] Meystre SM, Lovisb C, Bürklec T, Tognolad G, Budrionise A, Lehmann CU. Clinical data reuse or secondary use: Current status and potential future progress. In: IMIA Yearbook of Medical Informatics. Germany: IMIA and SchattauerGmbH; 2017. pp. 38-52. DOI: 10.15265/IY-2017-007

[47] VandeVusse A, Mueller J, Karcher S. Qualitative data sharing: Participant understanding, motivation, and consent. Qualitative Health Research. 2022;**32**(1):182-191. DOI: 10.177/10497323211054058

[48] Cumyn A, Dault R, Barton A, Cloutier A-M, Ethier J-F. Citizens, research ethics committee members and researchers' attitude toward information and consent for the secondary use of health data: Implications for research within learning health systems. Journal of Empirical Research on Human Research Ethics. 2021;**16**(3):165-178

[49] Tsai F-J, Junod V. Medical research using governments' health claims databases: With or without patients' consent? Journal of Public Health. 2018;**40**(4):71-877. DOI: 10.1093/pubmed/fdy034

[50] Mostert M, Bredenoord AL, Biesaart MCIH, van Delden JJM. Big Data in medical research and EU data protection law: Challenges to the consent or anonymise approach. European Journal of Human Genetics. 2016;**24**:956-960. DOI: 10.1038/ejhg.2015.239

[51] O'Neill O. Autonomy and Trust in Bioethics. United Kingdom: Cambridge University Press; 2002. p. 17

[52] Kerasidou A. Trust me, I'm a researcher!: The role of trust in biomedical research. Medicine, Health Care and Philosophy. 2017;**20**:43-50. DOI: 10.1007/s11019-016-9721-6

[53] Maseme M. Commodification of biomaterials and data when funding is contingent to transfer in biobank research. Medicine, Health Care and Philosophy. 2021:1-9. DOI: 10.1007/s11019-021-10042-3

[54] Wilkins CH. Effective engagement requires trust and being trustworthy. Medical Care. 2018;**56**(10):S6-S8

[55] Kraft S, Cho M, Gillespie K, Halley M, Varsava N, Ormond K, et al. Beyond consent: Building trusting relationships with diverse populations in precision medicine research. American

Journal of Bioethics. 2018;**18**(4):3-20. DOI: 10.1080/15265161.2018.1431322

[56] Moodley K, Singh S. "It's all about trust": Reflections of researchers on the complexity and controversy surrounding biobanking in South Africa. BMC Medical Ethics. 2016;**17**(57):1-9. DOI: 10.1186/s12910-016-0140-2

[57] Mcdonald M, Townsend A, Cox SM, Paterson ND, Lafrenière D. Trust in health research relationships: Accounts of human subjects. Journal of Empirical Research on Human Research Ethics. 2016:35-47. DOI: 10.1525/jer.2008.3.4.35

[58] McDavitt B, Bogart LM, Mutchler MG, Wagner GJ, Green HD Jr, Lawrence SJ, et al. Dissemination as dialogue: Building trust and sharing research findings through community engagement. Public Health, Research, Practice, and Policy. 2016;**13**:150473. DOI: 10.5888/pcd13.150473

[59] Platt J, Kardia S. Public trust in health information sharing: Implications for biobanking and electronic health record systems. Journal of Personalised Medicine. 2015;**5**:3-21. DOI: 10.3390/jpm5010003

[60] O'Brien J, Chantler C. Confidentiality and the duties of care. Journal of Medical Ethics. 2003;**29**:36-40. DOI: 10.1136/jme.29.1.36

[61] Molyneux S, Mulupi S, Mbaabu L, Marsh V. Benefits and payments for research participants: Experiences and views from a research Centre on the Kenyan coast. BMC Medical Ethics. 2012;**13**(13):1-15

[62] Allen J, McNamara B. Reconsidering the value of consent in biobank research. Bioethics. 2011;**25**(3):155-166. DOI: 10.1111/j.1467-8519.2009.01749.x

[63] Tai MC-T. Deception and informed consent in social, behavioral, and educational research (SBER). Tzu Chi Medical Journal. 2012;**24**:218-222. DOI: 10.1016/j.tcmj.2012.05.003

[64] Nicol D. Public Trust, intellectual property and human genetic databanks: The need to take benefit sharing seriously. JIBL. 2006;**3**:89-103

[65] Nicol D, Critchley C. Benefit sharing and biobanking in Australia. Public Understanding of Science. 2011;**21**(5):534-555. DOI: 10.1177/0963662511402425

[66] Johnsson L, Helgesson G, Hansson MG, Eriksson S. Adequate trust avails, mistaken trust matters: On the moral responsibility of doctors as proxies for patients' trust in biobank research. Bioethics. 2012;**1-8**. DOI: 10.1111/j.1467-8519.2012.01977.x

[67] Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: Understanding mistrust about research participation. Journal of Health Care for the Poor and Underserved. 2010;**21**(3):879-897. DOI: 10.1353/hpu.0.0323

[68] Durant RW, Legedza AT, Marcantonio ER, Freeman MB, Landon BE. Different types of distrust in clinical research among whites and African Americans. Journal of the National Medical Association. 2011;**103**(2):123-130

[69] Brandon DT, Isaac LA, LaVeist TA. The legacy of Tuskegee and Trust in Medical Care: Is Tuskegee responsible for race differences in mistrust of medical care? Journal of the National Medical Association. 2005;**97**(7):951-956

[70] Centers for Disease Control and Prevention. The Tuskegee Timeline. 2021. Available from:

https://www.cdc.gov/tuskegee/timeline.htm#:~:text=In%201932%2C%20the%20USPHS%2C%20working,Syphilis%20Study%20at%20Tuskegee%E2%80%9D. [Accessed: July 05, 2022]

[71] ScienceDirect. Tuskegee Syphilis Experiment. 2022. Available from: https://www.sciencedirect.com/topics/medicine-and-dentistry/tuskegee-syphilis-experiment/pdf. [Accessed: July 05, 2022]

[72] Silvers A. Historical vulnerability and special scrutiny: Precautions against discrimination in medical research. The American Journal of Bioethics. 2004;**4**(3):56-57

[73] Harvard Law School. Nuremberg Trials Project. 2020. Available from: https://nuremberg.law.harvard.edu/nmt_1_intro. [Accessed: July 05, 2022]

Chapter 6

Predictive Data Analysis Using Linear Regression and Random Forest

Julius Olufemi Ogunleye

Abstract

A statistical technique called predictive analysis (or analytics) makes use of machine learning and computers to find patterns in data and forecasts future actions. It is now preferred to go beyond descriptive analytics in order to learn whether training initiatives are effective and how they may be enhanced. Data from the past as well as the present can be used in predictive analysis to make predictions about what might occur in the future. Businesses can improve upcoming learning projects by taking actionable action after identifying the potential risks or possibilities. This chapter compares two predictive analysis models used in the predictive analysis of data: the Generalized Linear Model with Linear Regression (LR) and the Decision Trees with Random Forest (RF). With an RMSE (Root Mean Square Error) of 0.0264965 and an arithmetic mean for all errors of 0.016056967, Linear Regression did better in this analysis than Random Forest, which had an RMSE of 0.117875 and an arithmetic mean for all errors of 0.07062315. Through the hyper-parameter tuning procedure, these percentage errors can still be decreased. The combined strategy of combining LR and RF predictions, by averaging, nevertheless produced even more accurate predictions and will overcome the danger of over-fitting and producing incorrect predictions by individual algorithms, depending on the quality of data used for the training.

Keywords: data analysis, predictive data analysis, linear regression, random Forest, generalized linear model, decision trees

1. Introduction

Data analysis is the process of analyzing data to increase productivity and business growth. It involves steps like data cleansing, transformation, inspection, and modeling to perform market analysis, gather hidden data insights, enhance business studies, and generate reports based on the available data using tools like Tableau, Power BI, R and Python, Apache Spark, etc.

1.1 Predictive analysis

Predictive analytics also referred to as predictive analysis, is a subset of data analysis that focuses on creating future predictions from data. Other types of data analysis

exist, such as descriptive and diagnostic analysis, but predictive analysis is very well-liked in business analysis because it is crucial for making wise decisions. In any case, predictive analysis typically uses a variety of statistical models, techniques, and tools that all aid in understanding the patterns in datasets and making predictions. Data description and sorting are only a small part of predictive analysis. It largely relies on sophisticated models created to conclude from the data it encounters. To predict future trends, these models evaluate previous and present data using algorithms and machine learning. Depending on the particular requirements of people using predictive analysis, each model varies. Predictive analysis is very useful for assessing business decisions. This is because decisions effectively involve understanding their effects and basing them on projections of how a project, group, environment, or other entity will perform. A few typical fundamental models that are often used include:

- *Decision trees:* Use branching to illustrate the potential outcomes of each option or course of action.
- *Regression techniques:* Help in deciphering the connections between variables.
- *Neural networks:* Make use of algorithms to discover potential connections between data sets.

Prediction is a key component of data mining. Predictive analysis is a method for forecasting future patterns from current or historical data. As a result, businesses will be able to forecast future data trends. It can take many different forms, but some of the most advanced models make use of machine learning and artificial intelligence [1].

1.2 Models for predictive analysis

Predictive analysis encompasses several different types of data analysis models. Most of these are regression models, which aim to determine the connections between two or more variables. They can aid in predicting the value of an unknown variable as the value of a known variable changes by recognizing the links between these variables.

i. Generalized Linear Model - *Linear Regression*

The linear regression model is the most basic predictive analysis approach. In this approach, it is presumed that an unknown variable's value will scale linearly with a known variable's value. To track straightforward relationships and anticipate their future, such as expanding a customer base, linear regression models might be useful.

ii. Decision Trees - Random Forests

Random forests are machine learning models that, among other things, can be used to model regression. They are appropriate for huge data collections with several variables and are made up of some decision trees.

iii. Neural Networks

A cutting-edge tool for predictive analysis is neural networks. They are a collection of digital or biological neurons that talk to one another. A neural network changes shape and comes to new conclusions based on the data.

1.3 Predictive analysis tools

Aside from models, there are many specific tools available for conducting predictive analysis. These technologies aid in the discovery of connections that can be utilized to establish future predictions on data. They take on the bulk of the user's work by incorporating many statistical models used in the predictive analysis [2].

i. RapidMiner Studio

IBM provides a variety of predictive analytics technologies, including its premier SPSS Statistics software offering, as SaaS solutions. The system, which offers a variety of predictive analysis models, is primarily aimed at enterprise users.

ii. KNIME

Many of the functionalities of RapidMiner Studio are also available in the open-source data analysis tool KNIME. It appears to be made for more experienced users, though.

iii. IBM Predictive Analytics

A well-liked commercial tool for all types of predictive analysis is RapidMiner Studio. It aids in data collection, processing, and application of various statistical models to produce insightful results.

iv. SAP Predictive Analytics

SAP has a well-known SaaS product in the predictive analytics market. The developer of enterprise management software provides an analytics cloud for business users that is implemented similarly to IBM's.

2. Related works

2.1 Predictive analysis using linear regression with SAS (Bafna J., 2017)

According to Bafna J., a scalar dependent variable and one or more independent variables that are explanatory are connected using linear regression. The best-fitted straight line across the points in linear regression, one of the most widely used prediction methods, is referred to as a regression line. To demonstrate his thesis, the author gave the example of estimating people's weights based on their heights. The dependent variable in this situation is the weight, which needs to be predicted, and the independent variable is the height. The following outcomes were obtained using SAS' PROC REG to utilize linear regression to determine the relationship between two variables:

- The model has an R-squared score of 0.9541 (95.41%) > 0.7 (70%), suggesting that it suited the data well.

- With a P value of 0.00080.05, height was a significant variable in the model.

- To check for any outliers in the observations, the value of r was determined. If the value of r was greater than 2 or less than −2, the observations were considered outliers. (Note: $-2 < r < 2$.)

No observations deviated from the outliers range, leading the author to conclude that a major variable accounted for 95% of the person's weight (height) [3].

2.2 Random forest model to identify factors associated with anabolic-androgenic steroid use (Manoochehri Z., Barati M., Faradmal J. and Manoochehri S., 2021)

Androgenic-anabolic steroids are one form of doping bodybuilders frequently take (AAS). In addition to breaking athletic ethics, using AAS would harm one's physical and mental health. This study used a prototype willingness model to identify the key characteristics influencing AAS use among bodybuilders (PWM). A total of 280 male bodybuilders were chosen in 2016 utilizing multistage sampling from the bodybuilding clubs in Hamadan city for the analytical cross-sectional study. The data was then gathered through a self-administered questionnaire that included demographic data and PWM components, and a random forest model was also employed to evaluate the data. The most crucial elements in defining behavioral intention were behavioral willingness, attitude, and prior AAS usage. Additionally, BMI, attitude, subjective standards, and prototypes had the biggest impacts on predicting behavioral willingness to take AAS. Additionally, it was found that behavioral intention was more significant than behavioral willingness in predicting AAS usage. The findings indicate that, in comparison to the social reaction path, the reasoned action path has a stronger impact on predicting the use of AAS among bodybuilders [4].

2.3 Linear regression analysis study (Kumari K. and Yadav S., 2021)

According to the authors, linear regression is a statistical method for determining the value of a dependent variable based on an independent variable and determining the relationship between two variables. It is a modeling method in which one or more independent variables are used to forecast a dependent variable, and, according to the authors, it is the most widely applied statistical method. The chapter provided an overview of the underlying ideas and examples of performing linear regression calculations using SPSS and Excel (**Table 1**).

Regression statistics	Values	Explanation
Multiple R	0.96332715	Correlation coefficient: 1 means perfect correlation and 0 means none
R^2	0.927999198	Coefficient of determination: How many points fall on the regression line. Here, 92% points fall within the line
Adjusted R^2	0.891998797	Adjusted R^2: Adjusts for multiple variables, use with multiple variables
SE	516.3490153	
Observations	7	

Table 1.
Summary output.

According to the table above, multiple R is the correlation coefficient, where 1 (one) denoted a perfect correlation, and 0 (zero) denoted a lack of correlation. The factors might account for 92% of the variation according to the R Square coefficient of determination. Adjusted R-squared was utilized because it was corrected for many factors. The best methods for figuring out the link between two variables, according to the authors, were correlation and linear regression. Correlation measures the strength of a linear relationship between two variables, whereas regression describes the relationship as an equation. In the essay, straightforward examples using SPSS and Excel were provided to illustrate linear regression analysis and urge readers to adopt these techniques to analyze their data [5].

3. Methods

3.1 Linear regression

A machine learning technique called linear regression enables the conversion of numerical inputs into numerical outputs and the fitting of a line through the data points. In other words, a method of modeling the relationship between one or more variables is called linear regression (**Figure 1**) [6]. From a machine learning perspective, this is done to accomplish generalization, which enables the model to forecast results for inputs it has never seen before. It is one of the most well-known concepts in statistics and machine learning, and since it is so crucial, it consumes a sizable chunk of almost every Machine Learning course [7].

$$y = mx + c.$$

where x is the score of the independent variable, m is the regression coefficient, c is the constant, and x is the independent variable, is the formula for every straight line on a plot.

The formula for this in machine learning is h(x) = w0 + w1.x, where x is the input feature, w0 and w1 are weights, and h(x) is the label (i.e., y-value). The goal of linear regression is to identify the weights (w0 and w1) that produce the line that fits the input data the best (i.e. x features) [8].

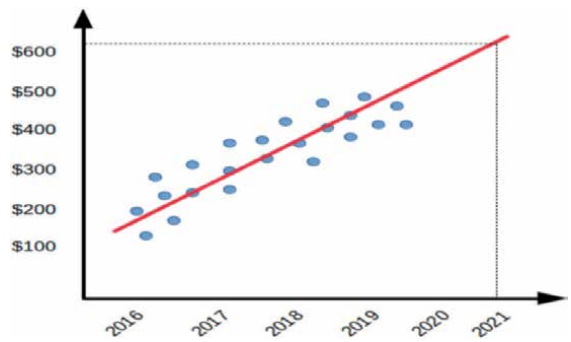


Figure 1.
Graphical illustration of a line (in red) generated by linear regression [3].

3.2 Random Forest

Machine learning methods for solving classification and regression issues include Random Forests. It uses ensemble learning, a method for solving complicated issues by combining a number of classifiers. The decision trees used in the random forest algorithm are numerous. The random forest algorithm creates a "forest" trained via bagging or bootstrap aggregation [9]. The accuracy of machine learning algorithms is increased by bagging, an ensemble meta-algorithm. Based on the predictions of the decision trees, the (random forest) algorithm determines the result. It makes predictions by averaging or averaging out the results from different trees [10]. The accuracy of the result grows as the number of trees increases. The decision tree algorithm's drawbacks are eliminated by a random forest, which also decreases dataset overfitting and boosts precision. Without requiring numerous configurations in packages, it generates forecasts (like Scikit-learn) [11].

3.3 The random Forest Algorithm's features

It overcomes the problem of overfitting in decision trees and is more accurate than the decision tree technique. In every random forest tree, a subset of characteristics is randomly chosen at the node's splitting point, providing an efficient approach of addressing missing data [12].

The fundamental distinction between the random forest method and the decision tree algorithm is that the latter randomly selects the root nodes and groups the nodes (**Figure 2**). To produce the necessary forecast, the random forest uses the bagging approach [13]. Bagging entails using multiple samples of data (training data) as opposed to a single sample. Predictions are made using features and observations from a training dataset. Depending on the training information employed by the random forest algorithm, the decision trees generate various results. The highest ranking of these outputs will be chosen as the final output [7].

```
proj.dtypes

Project_Id              int64
Project_Class           int64
TeamExp                 int64
ManagerExp              int64
YearEnd                 int64
Project_Length          int64
Actual_Effort           int64
Transactions_Count      int64
Entities                int64
Project_Size_AFP        int64
Adjustment_Points       int64
Project_Size_UFP        int64
Language_Used           int64
Productivity_Factor     int64
Effort_Calculation      int64
dtype: object
```

Figure 2.
Data types.

3.3.1 Advantages of random forest

- It is capable of both classification and regression tasks.
- A random forest generates accurate predictions that are simple to comprehend.
- It has efficient handling of big datasets.
- Compared to the decision tree method, the random forest algorithm is more accurate at predicting outcomes [14].

3.3.2 Disadvantages of random forest

- More resources are needed for calculation when utilizing a random forest.
- It takes longer than a decision tree approach [15].

4. Discussions

In this chapter, the predictive analysis methods Linear Regression and Random Forest are compared. Data on software cost estimation was obtained from Kaggle, and the database contained details on the function point-measured size of the implemented program. To determine which model had the lowest error and anticipated the software cost, H2O AutoML was used. The expected performance of machine learning systems may be greatly impacted by erroneous and noisy input. Poor data quality, notably the significant occurrence of missing values and outliers, may result in inconsistent and incorrect conclusions. Therefore, a key stage in developing ML models is pre-processing data through selection, cleaning, reduction, transformation, and feature selection (**Table 2**).

The project dataset was divided into the train (80%) and test (20%) halves for modeling purposes using H2O. The first one was used to create models, while the second one was used to verify their capacity to estimate effort. H2O AutoML was used to apply two data mining prediction methods (Generalized Linear Models - Linear Regression (LR) and Decision Trees - Random Forest (RF)) for both dependent variables. In order to assess their potential utility for implementation inside companies, error and accuracy measures were contrasted. The error measurements used to evaluate the accuracy of software estimate models were *Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE),* and *Root Mean Squared Log Error (RMSLE)* (**Table 3**).

Linear Regression outperformed Random Forest, as shown in the graph above. Additionally, there is very little variation between the models' RMSLE and MAE. Therefore, it can be concluded that there were no significant errors. The difference in prediction accuracy between the algorithms was essentially small, and each one could be employed independently for the examination of the predictions. In conclusion, both models are quite good at making predictions. However, in this instance, linear

Project_ Id	Project_ Class	TeamExp	ManagerExp	YearEnd	Project_ Length	Actual_ Effort	Transactions_ count	Entities	Project_ Size_AFP	Adjustment_ Function_ Points	Project_ Size_UFP	Language	Productivity_ Factor	Effort_ Calculatlon
1	1	1	4	85	12	5152	253	52	305	34	302	1	20	6211
2	2	0	0	86	4	5635	197	124	321	33	315	1	29	9182
3	3	4	4	85	1	805	40	60	100	18	83	1	20	2013
4	4	0	0	86	5	3829	200	119	319	30	303	1	19	6107
5	5	0	0	86	4	2149	140	94	234	24	208	1	15	3592
6	6	0	0	86	4	2821	97	89	186	38	192	1	29	5409
7	7	2	1	85	9	2569	119	42	161	25	145	2	22	3479
8	8	1	2	83	13	3913	186	52	238	25	214	1	21	5007
9	9	3	1	85	12	7854	172	88	260	30	247	1	46	11,872
10	10	3	4	83	4	2422	78	38	116	24	103	1	31	3602
11	11	4	1	84	21	4067	167	99	266	24	237	1	24	6478
12	12	2	1	84	17	9051	146	112	258	40	271	1	62	15,994
13	13	1	1	84	3	2282	33	72	105	19	88	1	69	7261
14	14	3	4	85	8	4172	162	61	223	32	216	1	26	5743
15	15	4	4	85	9	4977	223	121	344	28	320	1	22	7678
16	16	3	2	85	8	1617	119	48	167	26	152	2	14	2269
17	17	4	3	85	8	3192	57	43	100	43	108	1	56	5600
18	18	4	4	86	14	3437	68	316	384	20	326	2	51	19,409
19	19	3	4	87	14	4494	100	386	395	21	340	2	45	17,751
20	20	4	2	86	5	840	58	34	92	29	86	1	14	1332
21	21	4	4	86	12	14,973	318	269	587	34	581	2	47	27,639
22	22	2	4	85	18	5180	88	170	258	34	255	1	59	15,187

Project_ Id	Project_ Class	TeamExp	ManagerExp	YearEnd	Project_ Length	Actual_ Effort	Transactions_ count	Entities	Project_ Size_AFP	Adjustment_ Function_ Points	Project_ Size_UFP	Language	Productivity_ Factor	Effort_ Calculatlon
23	23	2	4	86	5	5775	306	132	438	37	447	1	19	8266
24	24	4	1	87	20	10,577	304	78	382	39	397	1	35	13,291
25	25	1	4	86	8	3983	89	200	289	33	283	1	45	12,934
26	26	4	1	85	14	3164	86	230	316	33	310	1	37	11,626
27	27	2	0	86	6	3542	71	235	306	37	312	1	50	15,266
28	28	3	1	85	14	4277	148	324	472	39	491	1	29	13,640
29	29	4	4	85	16	7252	116	170	286	27	263	1	63	17,880
30	30	4	1	85	14	3948	175	277	452	37	461	1	23	10,197
31	31	4	3	86	6	3927	79	128	207	37	190	1	50	10,790

Table 2.
Sample of sourced data.

Algorithm	RMSE	MSE	MAE	RMSLE	Arithmetic Mean
LR	0.026497	0.000702066	0.018321	0.018709	0.016056967
RF	0.117875	0.0138946	0.069672	0.081051	0.07062315

Table 3.
Errors and accuracy measures.

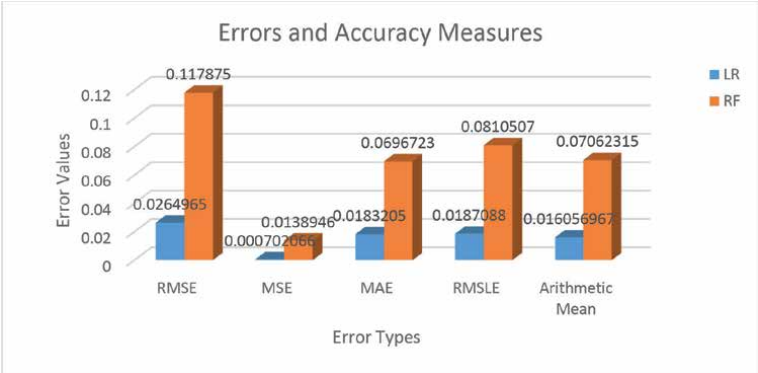


Figure 3.
Graphical representation of the errors and accuracy measures.

regression outperformed the other model. If deployed for a specific company and trained using a homogeneous dataset, models may be more accurate (**Figure 3**).

5. Conclusion

Almost every field uses predictive analysis, even though it has drawn some criticisms. With more information, future outcomes can be predicted with relative accuracy. This makes it possible for organizations and businesses to make educated decisions to increase production. Learning the methods of predictive analysis has become essential for jobs in data science and business analysis since it has numerous applications in every conceivable industry. In this investigation, Random Forest had an RMSE of 0.117875 and arithmetic mean for all errors of 0.07062315, while Linear Regression had an RMSE of 0.0264965 and arithmetic mean of 0.016056967. Through the hyper-parameter tuning procedure, these percentage mistakes can still be decreased.

This study compares the Generalized Linear Model with Linear Regression and the Decision Trees with Random Forest models for predictive analysis. Additionally, a merged strategy was investigated, which used the arithmetic mean to combine the predictions of the two models. The outcomes demonstrated that distinct data mining techniques might be applied to make predictions. The combined strategy of combining LR and RF predictions by averaging nevertheless produced even more accurate predictions and will overcome the danger of over-fitting and producing incorrect predictions by individual algorithms, depending on the quality of data used for the training. To maintain accuracy in a project's changing environment, it is important to remember that project management offices should ensure good input data quality and model updates.

Acknowledgements

I, Julius Olufemi Ogunleye (the author), would like to express my gratitude to Ass. Prof. Zdenka Prokopova and Ass. Prof. Petr Silhavy for their support and guidance in making this research work possible. This work was supported by the Faculty of Applied Informatics, Tomas Bata University in Zlín, under Projects IGA/CebiaTech/2022/001.

Author details

Julius Olufemi Ogunleye
Tomas Bata University in Zlin, Czech Republic

*Address all correspondence to: juliusolufemi@yahoo.com

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

[1] Data Mining Techniques: Algorithm, Methods & top Data Mining Tools. Software Testing Help; March 2020. Available from: https://www.softwaretestinghelp.com/data-mining-techniques/

[2] Steeneken F, Ackley D. A Complete Model of the Supermarket Business. BPTrends ■ January 2012

[3] Bafna J. Predictive Analysis Using Linear Regression With SAS. Big Data Zone – DZone; 2017

[4] Manoochehri Z, Barati M, Faradmal J, Manoochehri S. Random forest model to identify factors associated with anabolic-androgenic steroid use. BMC Sports Sci Med Rehabil. 2021;**13**(1):30

[5] Kumari K, Yadav S. Linear regression analysis study. Curriculum in Cardiology—Statistics. 2021;**4**:33-36

[6] Sumiran K. An overview of data mining techniques and their application in industrial engineering. Asian Journal of Applied Science and Technology. 2018;**2**:947-953

[7] Mehmed K. Data Mining – Concepts, Models, Methods, and Algorithms. Edition – 2, Illustrated Edition. Wiley; 2011. ISBN 1118029127, 9781118029121

[8] Varshini AGP, Kumari KA. Predictive analytics approaches for software effort estimation: A review. Indian Journal of Science and Technology. 2020;**13**:2094-2103

[9] Nassif AB et al. Software development effort estimation using regression fuzzy models. Computational Intelligence and Neuroscience. 2019;**2019**:8367214

[10] Azzeh MA, Nassif B, Banitaan S. Comparative analysis of soft computing techniques for predicting software effort based use case points. IET Software. 2018;**12**(1):19-29

[11] Dejaeger K et al. Data mining techniques for software effort estimation: A comparative study. IEEE Transactions on Software Engineering. 2012;**38**(2):375-397. DOI: 10.1109/TSE.2011.55

[12] Weiss GM, Davison BD. Data Mining. In: Bidgoli H, editor. Handbook of Technology Management. John Wiley and Sons; 2010

[13] Berson A et al. An Overview of Data Mining Techniques. (Excerpts from the book 'Building Data Mining Applications for CRM' by Alex Berson, Stephen Smith, and Kurt Thearling). McGraw-Hill; 2005

[14] Data Mining Techniques: Algorithm, Methods & top Data Mining Tools. Software Testing Help; April 2020. Available from: https://www.softwaretestinghelp.com/data-mining-techniques/

[15] Kushwaha DS, Misra AK. Software Test Effort Estimation. (ACM SIGSOFT Software Engineering Notes – Page 3). May 2008;**33**(3)

Chapter 7

Field Programmable Reconfigurable Mesh (FPRM)

Esti Stein and Yosi Ben Asher

Abstract

Many application areas demand increasing amounts of processing capabilities. FPGAs have been widely used for improving this performance. FPRM (Field Programmable Reconfigurable Mesh) is a technique we propose to improve FPGA performance. A Reconfigurable Mesh (RM) consists of a grid of Processing Elements that use dynamic reconfigurations to create varying bus segments between them. The RM can thus perform computations such as Sorting or Counting in a constant number of steps. It has long been speculated that the RM's dynamic reconfigurations should replace the FPGA's static reconfigurations. We show that the RM is capable of not only speeding up specific computations such as sorting or summing, but also of speeding up the evaluation of Boolean circuits (BCs), which is the main purpose of the FPGA. Our proposed RM algorithm can evaluate BCs without causing size blowup. Furthermore, tri-state switching elements can be used instead of PEs in a grid.

Keywords: FPGA, reconfigurable mesh, DNF, Boolean circuits, tri-state

1. Introduction

FPGAs are integrated circuits that form a matrix of conigurable logic units (CLUs)[1] connected via programmable routing interconnects. By downloading different routing configurations to the FPGA, any circuit $C(x_0, \ldots, x_{n-1})$ can be embedded and then executed/evaluated. After embedding the circuit's topology in the FPGA, the circuit is executed every time a new input is received. Due to the FPGA's routing interconnects and CLUs, this evaluation mode makes the FPGA relatively slow compared to ASICs.

Assuming the circuit $C(x_0, \ldots, x_{n-1})$ that has been discussed previously, we wish to examine the possibility of speeding up the evaluation of C by using a dynamic mode of reconfiguration rather than the above-mentioned FPGA mode. Essentially, we devised an algorithm that evaluates $C(x_0, \ldots, x_{n-1})$ faster than its depth [1] (the longest path from the root/output to any leaf/input)[2]. In a sequence of reconfiguration steps this algorithm: 1) Spans bus segments on different subsets of $\{x_0, \ldots, x_{n-1}\}$ in parallel; 2) Uses a single broadcast in each of these bus segments, and computes in parallel the AND/OR/COUNTING-1 s(counting the # of '1's) of each segment; 3)

[1] See appendix A for all acronyms and abbreviations.

[2] A preliminary version of the following algorithm and results was presented as a poster in [1].

Computes C in a fixed small number of steps regardless of C's depth based on the above computations. The above algorithm uses a platform based on Reconfigurable Mesh (RM) [2], which is a 2D grid of Processing Elements (PEs) that uses dynamic bus reconfiguration to create varying bus segments for fast communication. Consequently, computations such as summation and sorting can be expedited.

Reconfigurable Mesh (RM) has been demonstrated to be able to perform parallel computations faster than the Parallel random access machine model (PRAM) [3], which is an abstract model for parallel computation. This includes $O(1)$ summing [4], $O(\log)$ integer summing [5], $O(1)$ multiplication [6], sorting [7], convex hall [8], graph algorithms [9, 10] and image processing [11]. Despite this potential power of the RM model, it has not yet been fully realized since the model assumes a signal can be transmitted along a bus/connected component in a single step regardless of the number of switches/ports. From this perspective, a variety of restricted RMs have been proposed. These include the RMBM [12], where only the structure of the RM's switch has been simplified but still busses with a linear number of switches are used. The SRGA [13, 14] proposed a mesh, where each row/column has a complete binary tree of reconfigurable switches, allowing to route messages between the leaves of this tree. [15] proposes a linear RM (LR-Mesh) bending cost, where the delay of a bus varies as a function of how many times it bends between rows and columns. It showed that for busses with a reasonable delay of at most $D = N^{\varepsilon}$ bends they can simulate algorithms for LR-Meshes in constant time. A bus of length $d(n) = n^{1/k}$ was suggested in [16], also showing that restricted RM algorithms can be directly coded in Verilog. This way of programming RM-algorithms overcomes most of the drawbacks of the C-like programming style proposed so far for RM-algorithms (e.g.,ARMlang [17]). However, they only addressed the problem of COUNTING-1 s. Our method of evaluating the circuit is partially based on the solution for COUNTING-1 s. [18] shows that integrating branching program with Boolean circuits is better than using each of them separately. Other realizations of the RM [19] were mainly to a small-size grid of Soft-CPUs and cannot be synthesized for large values of n.

A number of dynamic reconfiguration (DR) FPGAs have also been proposed, mainly for the purpose of speed acceleration. However, the main challenge was the reconfiguration delay. The use of DR is therefore rare [20]. There is also a method of addressing this problem, and it is commonly referred to as partial reconfiguration (PR) at runtime. PR can be implemented through external FPGA interfaces as well as special internal interfaces such as the ICAP on Xilinx devices [21]. Even so, PR is still primarily an auxiliary feature in modern commercial FPGAs rather than something with which the architecture is designed [22, 23]. Thus, PR design involves many details related to low-level architecture that require a high level of expertise. [24] proposed time-multiplexed DRFPGA, where registers are added to store computational states and partial results. Yet, only a few contexts are allowed because of area overhead. Memristors (RRAM), have also been applied as a programmable switch, as they are naturally more delay-efficient and lead to higher-performance FPGA architectures. However, [25, 26] only focus on the architectural repercussions of this technology. Very limited works investigate realistic RRAM-based circuit design constraints, while these have a strong impact on the final architectural performances. Fine-grain DR (FDR), described in [27], consists of homogeneous reconfigurable logic elements (LEs). It is possible to configure each LE as either a lookup table (LUT) or as an interconnect, or even as a combination of both. While this improves flexibility for allocating hardware resources between LUTs and interconnects, it still consumes a large amount of space. At first glance, this model seems to be close to what we have

proposed, but one of the main difference lies in the algorithm for evaluating the Boolean Circuit $C(x_0, \ldots, x_{n-1})$ faster than its depth.

Our approach to the speed-up evaluation problem is to use the Reconfigurable Mesh (RM). We propose an infrastructure called FPRM (Field Programmable Reconfigurable Mesh) which is a sub-model of the RM model based on current CMOS technology and adapted to the proposed algorithm. The FPRM consists of two-dimensional grids of switches $pe_{i,j}$, with each switch connected to four neighbors $pe_{i-1,j}$,$pe_{i+1,j}$,$pe_{i,j-1}$,$pe_{i,j+1}$ via four links. It allows reconfiguration of its internal links in different reconfiguration modes S_0,S_1,S_2, … as depicted in **Figure 1** (upper left). Each $pe_{i,j}$ has four registers used to read/write to each of the four links: *Nr* to read/write to the link connecting $pe_{i,j}$ to $pe_{i-1,j}$ and $Sr/Wr/Er$ to read/write to $pe_{i+1,j}/pe_{i,j-1}/pe_{i,j+1}$ respectively. Each $pe_{i,j}$ executes a program based on its current state, its coordinates *i*,*j* and the values of *Nr*,*Sr*,*Wr*,*Er*. Upon execution, each $pe_{i,j}$ can change its reconfiguration mode, its state, and the content of its registers *Nr*,*Sr*,*Wr*,*Er*. **Figure 1** contains a four instructions program (bottom left side) for executing COUNTING-1 s of a four bits input. As depicted in **Figure 1** right side, the execution of this program creates a bus whose bendings corresponds to the ′1′s input values. By examining the exit point (row number) of a signal sent through $S>$ we obtain the number of $1-bits$ in the input.

The second step of the FPRM computation is shown in **Figure 2** computing the DNF (Disjunctive Normal Form): $((x_0 \wedge x_1 \wedge x_2) \vee (\overline{x}_0 \wedge \overline{x}_3) \vee (\overline{x}_1 \wedge \overline{x}_3) \vee (x_2 \wedge \overline{x}_3))$ where each and-term (minterm) is computed in a different row. As with COUNTING-1 s, we broadcast the input values along the columns in the first step. If $pe_{i,j}$ is associated with $\wedge x_i \wedge \ldots$ in an and-term (minterm) and the input $x_i == 1$ then $pe_{i,j}$ switches to a connect mode selecting S_2, alternatively on $(x_i == 0)$ it switches to a disconnect mode selecting S_4. The opposite is performed if $pe_{i,j}$ is associated with $\wedge \overline{x}_i \wedge \ldots$. A *true* signal is sent from $S>$ for every row, while each disconnected $pe_{i,j}$ broadcasts a *false* signal from its *Er*. The or-term of these and-terms is computed in another broadcast along the last column. Obviously, the FPRM can be used to execute the $O(1)$ RM algorithms such as summing of n numbers, multiplication [6, 28], sorting, convex hull [2], graph algorithms [9] and image processing [11]). However, here we consider the problem of parallel evaluation of circuits with large depths for which no previous RM algorithm exists. Preliminary results demonstrate the FPRM feasibility and that it is likely to outperform FPGAs.

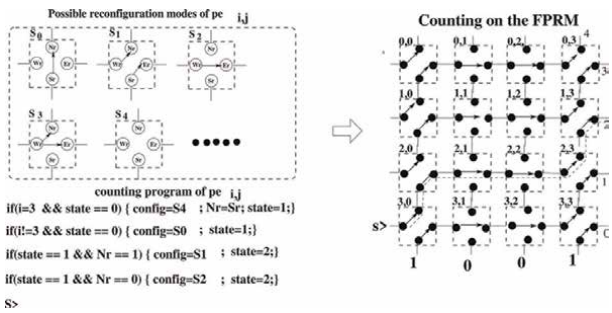


Figure 1.
The FPRM switches and a program to compute COUNTING-1 s using a 4×4 *FPRM.*

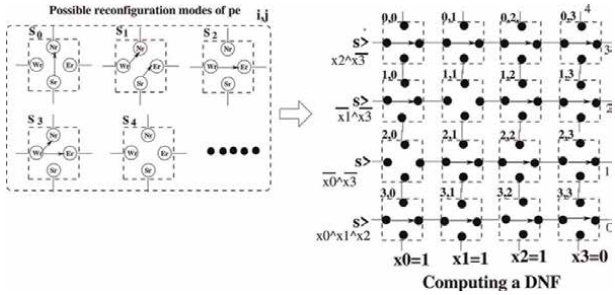


Figure 2.
Computing a DNF formula using a 4×4 *FPRM.*

According to the proposed algorithm, boolean circuits $C(x_0, \ldots, x_{n-1})$ can be evaluated in a constant number of FPRM steps regardless of C's depth. During compilation, we calculate the minimized DNF formula dnf_y for every possible result of COUNTING-1 s, which is $y = \sum_0^{n-1} x_i$. An FPRM program is generated for each of the DNFs (dnf_y) using the algorithm described in **Figure 2**. Each of these DNFs (dnf_y) is compiled into an FPRM program working in a similar manner to the algorithm described in **Figure 2**. At run-time, after performing COUNTING-1 s of the input, the FPRM selects the DNF-program dnf_y for y and executes it. Thus, in a constant number of steps this algorithm computes $C(x_0, \ldots, x_{n-1})$, using dymamic reconfiguration (DR). By first applying COUNTING-1 s, we get that the size of the FPRM grid needed to execute each of the $dnf_{y=0,\ldots,n-1}$ is less or equal to the size of the original $C(x_0, \ldots, x_{n-1})$. This is expected since each dnf_y in $C(x_0, \ldots, x_{n-1})$ is restricted to the case where $y = \sum_0^{n-1} x_i$. Further, the and-terms of dnf_y are packed in a 2D-FPRM layout with multiple and-terms computed in a single row (unlike **Figure 2**, where each and-term is computed separately).

The rest of the chapter is organized as follows. The following section describes the use of COUNTING-1 s operation in order to reduce the formula size. The next step describes the problem of fitting as many and-terms as possible into an FPRM grid, which is one of the most challenging aspects of the technique. The results of the experiments will be presented next, followed by a summary of the conclusions.

2. Using the counting-1 s operation to reduce formula size

For every possible outcome of $y = \sum_0{}^* n - 1 x_i$, the proposed algorithm starts by obtaining the minimized DNF formula, dnf_y. The input is divided into k segments containing *fracnk* bits, and the number of 1-bits is counted for each segment. For each of the $\left(\frac{n}{k} + 1\right)^k$ possible COUNTING-1 s results $y_1, \ldots, y_k$ $\left(y_i = 0 \ldots \frac{n}{k}\right)$, we compute a minimized DNF $dnf_{y_1,\ldots,y_k}$ by:

1. Building a truth table $T_{y_1,\ldots,y_k}$ of $C(x_0, \ldots, x_{n-1})$ for all the binary numbers $x_0, \ldots, x_{n-1}$ with y_i '1's in the i segment.

2. Our initial DNF is formed by the nonzero entries of $T_{y_1,\ldots,y_k}$, which we simplify using the Logic Friday Espresso package [29].

3. Using reduced DNF, dnf_y, monochromatic rectangles $M \times V$ are created by searching for all minterms in M, and variables in V in such a way that any minterm in M contains all variables in V. We obtain a smaller version of dnf_y by replacing each variable in M by a new variable (which is the AND of all variables in M). A separate step needs to be performed in order to compute the value of the new variables corresponding to monochromatic rectangles.

Consider the truth table T of the address-function of $n = 6$ boolean variables given in **Figure 3**.

$$F(a,b,c,d,e,f) = \begin{cases} c & <a,b> = 0,0 \\ d & <a,b> = 0,1 \\ e & <a,b> = 1,0 \\ f & <a,b> = 1,1 \end{cases}$$

T is arranged by COUNTING-1 s in $<a,b,c>$ $(y_1(a,b,c) \in \{0,1,2,3\})$, and COUNTING-1 s in $<d,e,f>$ $(y_2(d,e,f) \in \{0,1,2,3\})$. Since the address function has a very small formula to begin with

$$F(a,b,c,d,e,f) = ca'b' + da'b + eab' + fab$$

(where x' *is* $\neg x$), it is not expected that using COUNTING-1 s can significantly reduce the size of the remaining circuits $C^{y_1(a,b,c)=i, y_2(d,e,f)=j}(a,b,c,d,e,f)$. Indeed, the results in **Figure 3** shows that the minimal boolean formula for $C^{y_1(a,b,c)=2, y_2(d,e,f)=1}(a,b,c,d,e,f)$ is $a'bcde'f' + ab'cd'ef' + abc'd'e'f$ which is even larger than the original formula for the whole function $ca'b' + da'b + eab' + fab$.

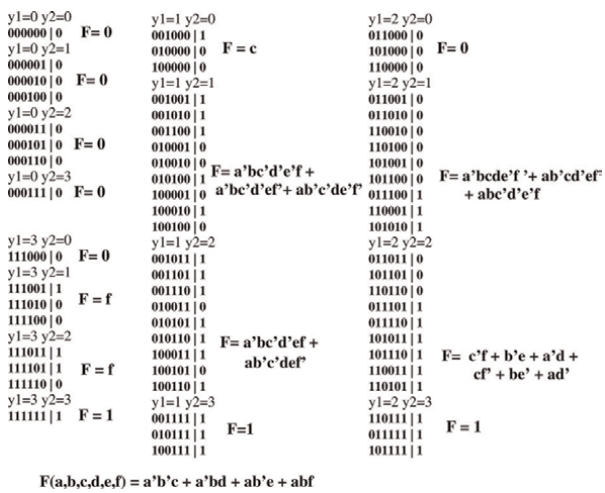


Figure 3.
Truth table of the address function arranged by COUNTING-1 s results.

However, this happens only for four out of the sixteen possible cases of $y_1(a, b, c) = i, y_2(d, e, f) = j$. In all the remaining 12 cases the boolean formula has one or no variables.

Yet, COUNTING-1 s is very helpful for the multiplication function

$$F(a, b, c, d, e, f) = 1 \ \ iff \ \ (a \cdot 2 + b) \cdot (c \cdot 2 + d)) \ mod \ 4 = (e \cdot 2 + f)$$

The results depicted in **Figure 4** shows that in all sixteen cases, the minimal boolean formula for $C^{y_1(a,\ b,\ c)=i, y_2(d,\ e,\ f)=j}(a, b, c, d, e, f)$ is very small.

Figure 5 depicts the largest dnf_y for $C = STCON(x_0, \ \dots, x_{48})$ of a seven nodes directed graph where the input is 7×7 adjacency matrix of the graph. In this case we selected $k = \sqrt{49} = 7$, hence $y = <y_1, \dots, y_7> \quad y_i = 0 \dots 7$. Out of all the COUNTING-1 s cases for $STCON(x_0, \ \dots, x_{48})$, **Figure 5** depicts the worst/largest dnf_y obtained. The dnf_y of **Figure 5** should be read as follows:

- *stcon e n* means STCON for graph of size n, while e is the number of entries in the adjacency matrix of the graph.
- *.i e* is the number of variables denoted by e (a variable for every entry in the adjacency matrix).
- *.p m* where m denotes the number of rows where the function is evaluated to TRUE.
- The rows are composed of e variables $a_0, \dots a_{e-1}$, where $-/1/0$ denotes $don't - care/a_i/a_{i'}$.

y0=0,y1=1		y0=1,y1=2		y0=2,y1=2	
000001 | 0	F=0	001011 | 0		011011 | 0	
y0=0,y1=0		001101 | 0		011101 | 0	
000000 | 1	F=1	001110 | 0		011110 | 0	
y0=0,y1=1		010011 | 0	F= be'	101011 | 0	
000010 | 0	F=e	010101 | 1		101101 | 0	F = f'b'
000100 | 1		010110 | 0		101110 | 1	
y0=0,y1=2		100011 | 0		110011 | 0	
000011 | 0		100101 | 0		110101 | 0	
000101 | 0	F=0	100110 | 1		110110 | 0	
000110 | 0		y0=1,y1=3		y0=2,y1=3	
y0=0,y1=3		001111 | 0		011111 | 1	
000111 | 0	F=0	010111 | 0	F = 0	101111 | 0	F = b
y0=1,y1=0		100111 | 0		110111 | 1	
001000 | 1		y0=2,y1=0		y0=3,y1=0	
010000 | 1	F=1	011000 | 0		111000 | 0	F = 0
100000 | 1		101000 | 1	F = a	y0=3,y1=1	
y0=1,y1=1		110000 | 1		111001 | 0	
001001 | 0		y0=2,y1=1		111010 | 1	F = e
001010 | 0		011001 | 0		111100 | 0	
001100 | 1		011010 | 1		y0=3,y1=2	
010001 | 0		011100 | 0	F = ea'	111011 | 0	
010010 | 0	F= cd	101001 | 0		111101 | 1	F = e'
010100 | 0		101010 | 0		111110 | 0	
100001 | 0		101100 | 0		y0=3,y1=3	
100010 | 0		110001 | 0		111111 | 0	F = 0
100100 | 0		110010 | 0			
		110100 | 0			

Figure 4.
Truth table of the mult function arranged by COUNTING-1 s results.

```
#stcon 49 7 .i 49 .o 1 .p 16
-------------1---------------------------------- 1
------11------------------------------------- 1
------1--1-----00000---------------------------- 1
---------1----000000---------------------------- 1
------1---1-----------00000---------------------- 1
----------1----------000000---------------------- 1
--1----1------000000---------------------------- 1
---1---1-------------000000---------------------- 1
------1---1----00000--0-000---------------------- 1
------1--1-----00-00--00000---------------------- 1
--1-------1---000000--0-000---------------------- 1
----------1---000000-00-000---------------------- 1
---1-----1-----00-00-000000---------------------- 1
---------1----000-00-000000---------------------- 1
---1---1------000000-00-000---------------------- 1
--1----1------000-00-000000---------------------- 1
.e
```

Figure 5.
Resulting formula for worst COUNTING-1 s case of 7×7 *STCON.*

As shown in **Figure 5**, the dnf_y contains only 16 minterms (and-terms), each containing 10 variables (or negation of variables) omitting some $all\ don't - care$ columns at the end of each row. To compare, we obtained the full circuit for $STCON(x_0, \ldots, x_{48})$ using VIVADO-HLS on the following C-code:

```
#define SQM 7
  for(k = 0; k < SQM; k++){
   #pragma HLS unroll factor = 7
    for(i = 0; i < SQM; i++){
     #pragm HLS unroll factor = 7
      for(j = 0; j < SQM; j++){
       #pragma HLS unroll factor = 7
        if (mat[i][j] || (mat[i][k] && 
            mat[k][j]))  mat[i][j] = 1; }
    }
   }
return(mat[1][SQM - 1]);
```

The synthesys results of the Verilog code obtained for the above code are: $clock - latency = 6ns$, $\#registers = 591$, $\#LUTs = 742$ and $\#MUXes = 782$. This is significantly larger than dnf_y of **Figure 5** which is a DNF with 143 gates and clock latency of less than $1ns$. The experiments show that using a fast pre-computation function of the inputs (COUNTING-1 s) can significantly reduce the size of $\max_y |dnf_y|$ compare to the size of the complete circuit.

The above dnf_y could have been further simplified by replacing monochromatic rectangles with new variables. As illustrated in the **Figure 6**, it is possible to simplify the largest dnf_y of a five node graph by replacing eight monochromatic rectangles (left-side) with new variables. The result is a reduction from $(94/14)/simeq6$ to $(49/14)/simeq3$ in the average number of variables per minterm. As each rectangle is evaluated at runtime, its subset of variables is logically ANDed. This can be achieved with a sub-bus of the FPRM that allows a zero-x_i to broadcast, on the value of 0.

3. Computing the FPRM layout of $\mathbf{dnf_y}$

Using the algorithm of **Figure 2**, we can evaluate the dnf_ys on the FPRM in three steps. Broadcasts representing the minterms of dnf_y are used to evaluate the DNF. As a

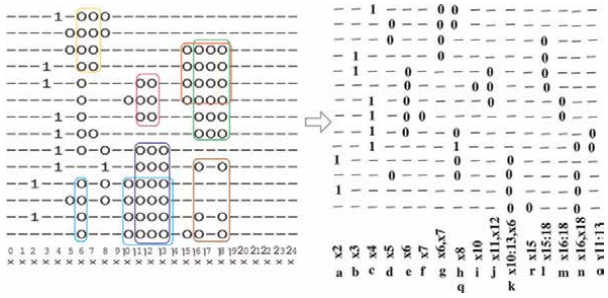


Figure 6.
Reducing the DNF size of a dnf_y by pre-computing monochromatic rectangles.

result, we can evaluate the DNF using the $n \times k$ grid of sub-FPRM, where n is the number of rows (minterms), and k is the number of variables (14×17 for the DNF shown in **Figure 6** right). There are two steps involved: broadcasting the values of each x_i over the columns in the sub-FPRM; and configuring each row as a single bus and computing the logical AND on each row. We can, however, pack several minterms/bus segments in one row, reducing the size of the FPRM sub-grid needed for the computation. Our 2D layout of a dnf_y can be optimized by swapping minterms in each level and arranging the literals in each minterm (node).

Figure 7 illustrates the optimized (by hand) layout of the DNF of **Figure 5** (called *LG*), wherein the minterms are arranged in six levels, each containing 1–4 minterms. Straight busses are used to broadcast the values of the literals in this layout. According to **Figure 7**, this layout also includes the extra duplications of $'b'$ and $'n'$ required. There is a significant improvement in the total area and max-switching length when compared to the simple method of arranging all minterms in one column. The optimized (by-hand) layout of **Figure 7** is also better compared to that of **Figure 5** when

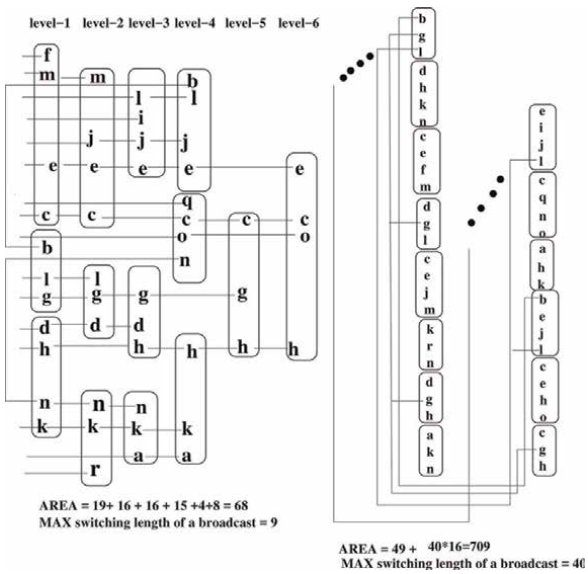


Figure 7.
Optimized FPRM layout.

used as an *LG*. In the following sections, we describe the details of the proposed algorithm to find a minimized *LG*.

3.1 First stage: find the level arrangement of minterms in the final FPRM layout

1. We build an intersection graph G^0 where each node corresponds to one minterm and each edge corresponds to an intersection between two such minterms. **Figure 8** (left) illustrates this graph for the DNF of **Figure 6**. The edges are labeled by the intersection size.

2. Next, we compute a maximal independent set (MIS) in G^0 that also maximizes the highest label edge in each of its nodes (as depicted in **Figure 8**). This MIS will be used as the first level in the FPRM layout we seek to build. Note that all the minterms in this MIS have no intersection in their variables, thus can be safely mapped to the same level of the layout.

3. We remove the MIS from G^0 creating G^1 and as depicted in **Figure 9**. compute the next MIS in G^1 creating the next level in the FPRM layout. This process of extracting MIS, forming the next level of minterms in the layout is repeated until there are no more MISs. **Figure 10** depicts the last step of this process.

3.2 Second stage: rearranging the minterms in each level

The position/index of each node/minterm in the final layout is computed as follows:

- Create a leveled graph *LG* whose nodes $V_{level,index}$ correspond to the minterms in each column of the layout previously obtained.

Figure 11 depicts the resulting *LG*.

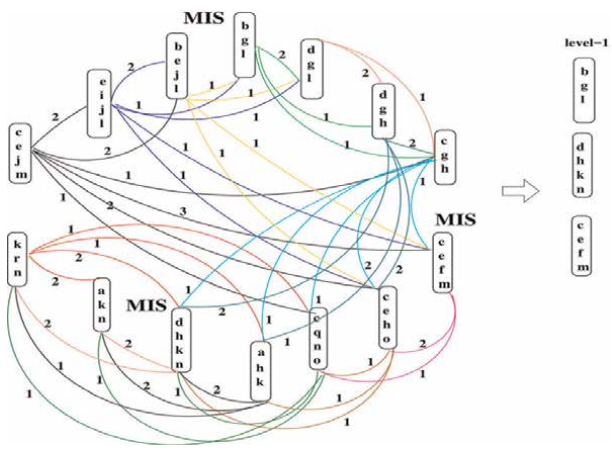


Figure 8.
The intersection graph G° and extracting first MIS.

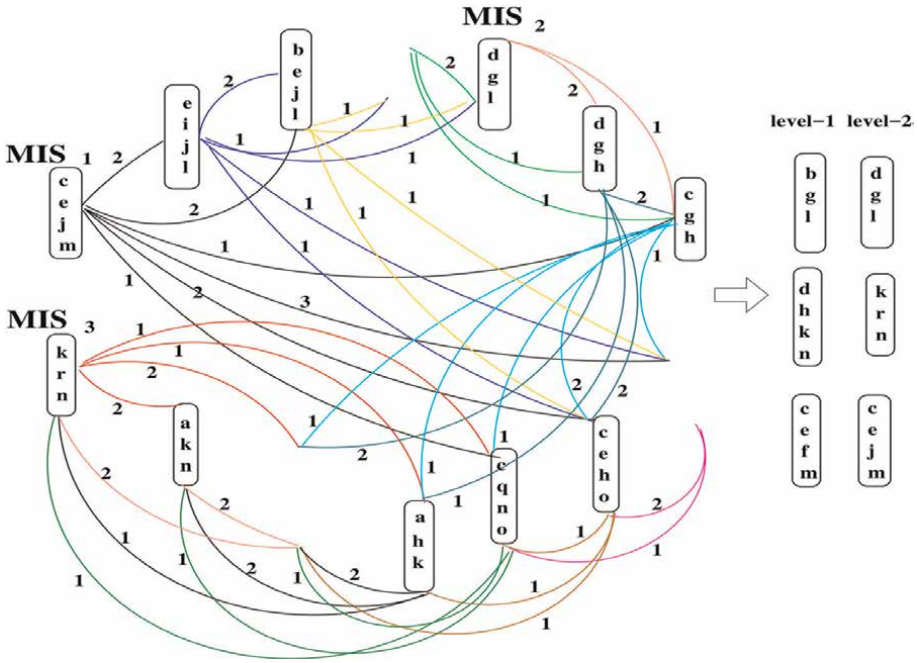


Figure 9.
G^2 and extracting the next level in the layout.

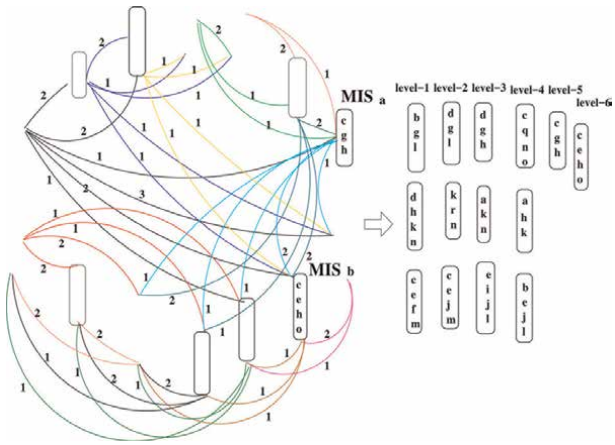


Figure 10.
G^5 and extracting the last level in the layout.

- Rearrange the minterms in each level of *LG* by: Finding a set of nodes (called "mid-cut"), one from each level of *LG*, and a partition of the remaining nodes in each level into a "left-part" and a "right-part" such that:
    - The number of edges between the left-part and the right-part is minimal.
    - The number of nodes in the left-part in each level is about the same as the number of nodes in the right-part of that level.

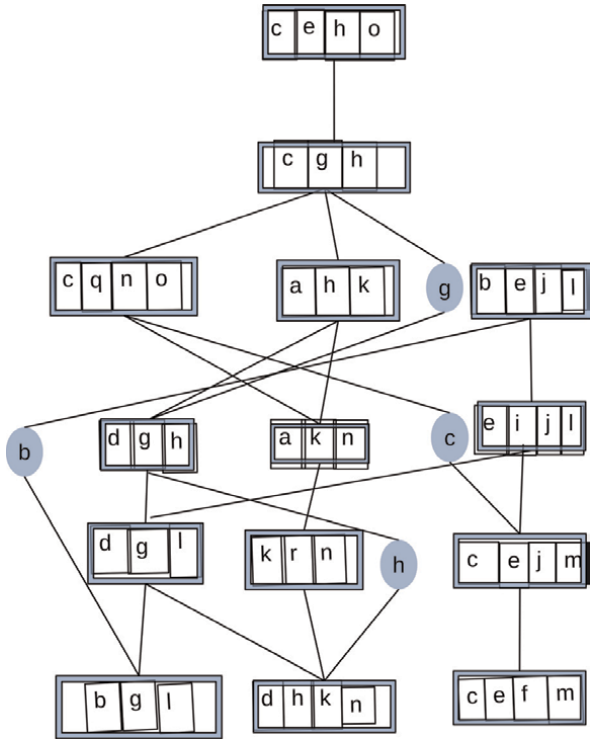


Figure 11.
The leveled graph (LG) of the MIS arrangment.

Figure 12 depicts finding a mid-cut and the resulting partition to a left-part and right-part. As can be seen, only one edge (the $'b'$) crosses from the left-part to the right-part in **Figure 12**. In the final FPRM layout, the mid-cut minterms will be stacked one on top of the other. Recursively, the process is applied to the left-part and the right-part until all nodes of *LG* are arranged in 2D.

3.3 Third stage: rearranging the order of literals in each minterm

1. For the current order of literals in each minterm, we expand the edges of the current *LG* to show the duplication of each literal from one level to another. We also add source-edges (arrows in **Figure 13**) depicting that the source of each literal comes from the lowest level below the present 2D layout of the *LG*. **Figure 13** illustrates the resulting graph called the literals graph *TG*.

2. Crossing edges between minterms that are in the same index at their level are eliminated by rearranging the literals inside one minterm.

3. Next we eliminate crossing edges by:

   - Replacing one pair of crossing edges with a down-going source-edge.

   - Reordering literals in the two minterms of the crossing edge. Crossing edges between minterms with the same level-index are resolved by rearranging the literals in those minterms.

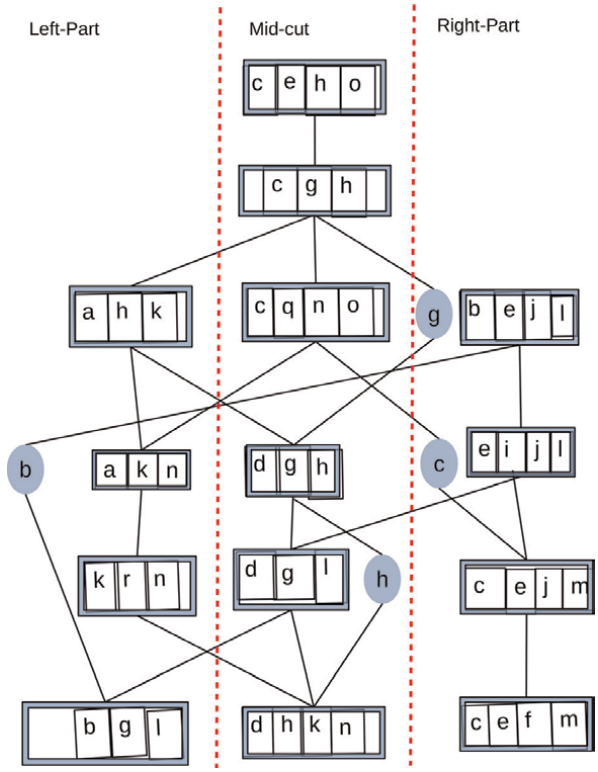


Figure 12.
Rearanging the minterms in LG's levels via mid-cuts.

The next edge selected to be replaced by a source edge is the one with the maximal number of crossings. For example, in **Figure 13** the first edge to be replaced by a source edge is the edge connecting the left $'b'$ in the first level to the right $'b'$ in the fourth level, as this edge cut acrosses seven edges. The process is repeated until there are no crossing edges, as shown in **Figure 14**. Since source edges will be aligned vertically later, crossing with source edges is not counted.

3.4 Fourth stage: completing the alignment

At this stage, the minterms in each level and the literals in each minterm have been arranged so that no crossing edges exist. Aligning the literals such that all the edges form straight vertical columns leads to the final FPRM layout:

1. We start by placing the bottom leftmost literal/source-edge at the bottom leftmost corner of the FPRM layout. We align all the edges connected to the left corner at a vertical "duplication" column (for duplicating literals, if needed).

2. Let v_i be the nearest node to the last aligned duplication column. We align v_i and the literal/source edges connected to it, into an adjacent vertical duplication column.

3. This is repeated until all edges are aligned into adjacent duplication columns.

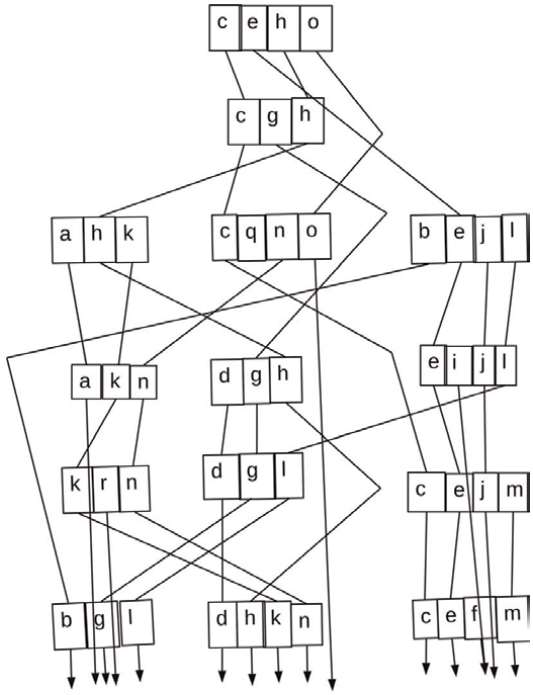


Figure 13.
Expanding current level graph LG to a literal graph TG.

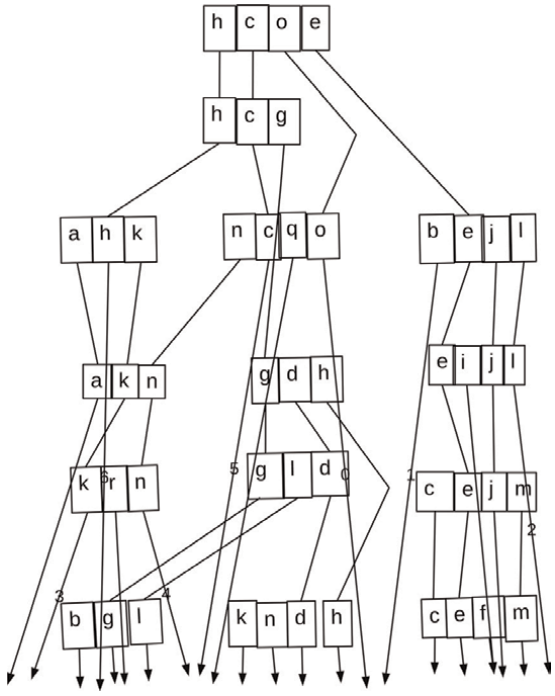


Figure 14.
Final arrangement of literals in each minterm.

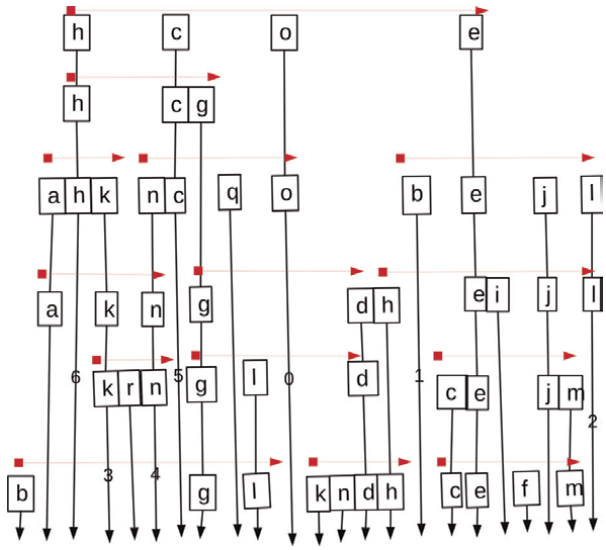


Figure 15.
Completing the alignment.

Figure 15 illustrates the resulting layout with a total area of about 143, which includes the area for broadcasting the duplicated literals (c,h,b,k,n,l).

Since the FPRMs constructed with literals as control entries into the tristate switches.

4. Realization and results

Tri-state switch is the natural candidate for the infrastructure logic realization of the final FPRM layout, as other switching devices such as NMOS are not supported by ASIC synthesis tools. For control input of '1', the output of the tri-state is exactly identical to its input, but for control input of '0' the output is high impedance (disconnected or 'z'). As a result, multiple tri-states can share the same output wires. The DNF is simply a $\vee_{i=1}^{n} m_i$, where m_i represent a minterm $m_i = \wedge_{j=1}^{|m_i|} a_j$, and a_j is a literal. For each literal, a minterm can be represented conceptually by a list of connected tri-states. The leftmost tri-state outputs '1' or 'z' based on input of '1' and control from the literal. The next tri-state, produces the input according to the next literal value in m_i, thus performs the operation of $\wedge_{j=1}^{|m_i|} a_j$. Since the output of each minterm m_i is '1' or 'z', their output wires can be connected directly, performing $\vee_{i=1}^{n} m_i$. **Figure 16** illustrates the FPRM for the DNF of two minterms $M = (a1 \wedge a2) \vee (a2 \wedge a3)$, where the literal values are represented by thick vertical lines (dark-gray for '1' and light-gray for '0'). Each (potential) literal a_j in a minterm m_i residing in row r consists of 6 tri-states and one encoder (depicted in the dashed line rectangle). The $cn_{r,j}$ tri-state is connecting the literal value to minterm m_i, providing that a_j appears in m_i. Otherwise, $cn_{r,j}$ will pass the incoming signal to the next literal. The output of $cn_{r,j}$ is the control of $tl_{r,j}$, transferring the input from a_{j-1} or disconnecting (producing 'z') given the value of a_j. According to the encoder's $e_{r,j}$ input, $tr_{r,j}$ will pass/hold the current signal to a_{j+1}, or $ci_{r,j}$ will start a new minterm

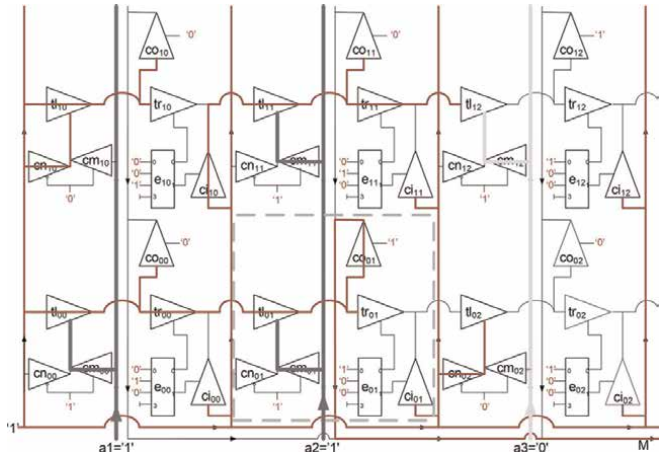


Figure 16.
Tri-state configuration of the FPRM.

calculation, sending '1' to the next literal on the right. Finally, the role of $co_{r,j}$ is to send down the output of the current minterm, providing that it is the rightmost literal in the current minterm. Note that the outputs of the minterms are connected together since these are outputs of tri-states ('1' or 'z'). The output of the DNF is indicated by M on the right-bottom side.

Based on the description in [30], the FPRM implementation is compared to an Island FPGA routing architecture. **Figure 17** shows a variant that contains: A logic unit with and/or-gate that is connected to a grid of $4 - bits\ N\ vertical\ buses \times 4 - bits\ N\ horizontal\ buses$ via two connection units. Any vertical-bus can be connected to any horizontal-bus using a crossbar-like routing unit. Additionally, vertical/horizontal busses can be disconnected, so that bends will not consume the entire bus.

All connections/disconnections and fuse operations are made by a back-to-back pair of tri-state devices, allowing bi-directional signals. ASIC synthesis results obtained with Synopsys Design compiler using a 160 nm cell library. As shown in **Table 1**, the FPRM architecture is 4X faster and more efficient in both power and

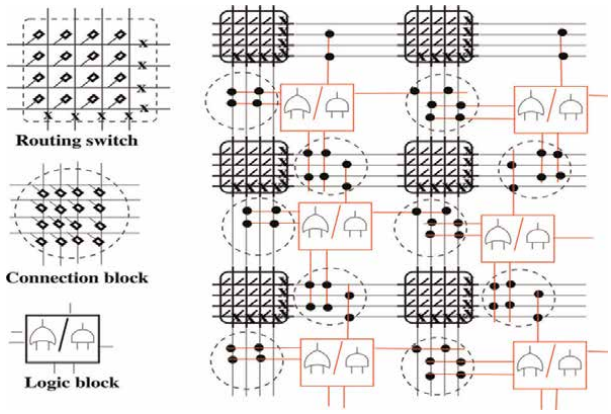


Figure 17.
The FPGA routing architecture used for the experiments.

Size	Area		Power		Clock latency	
	FPRM	FPGA	FPRM	FPGA	FPRM	FPGA
16 × 16	6156 uC	38,256 uC	66 uW	161 uW	6.98 ns	14.36 ns
12 × 12	3463 uC	21,528 uC	52 uW	149 uW	4.75 ns	11.27 ns
8 × 8	1221 uC	9579 uC	34 uW	148 uW	2.85 ns	8.18 ns
4 × 4	333 uC	2411 uC	12 uW	80 uW	1.36 ns	5.09 ns

Table 1.
Synthesis results comparing the FPRM vs. the FPGA routing infrastructure.

area[3] than the FPGA routing infrastructure. Based on the FPRM of **Figure 16** and the assumption that the counting stage requires two cycles. When the expected latency of the FPGA is added, we get about twice as fast performance from the FPRM.

A chain of switches (tri-states) selects whether values should be passed on or not in the circuit we designed. This idea will obviously work faster than a chain of *and* and *or* gates as implemented in FPGA. The fact that there are no switches along the wire that ends with M further accelerates the speed of receiving the output. This is triggered when a value of 1 comes out from one of the minterms. Given that the tri-state consumes power as an ordinary buffer, and the *and*/*or* operations ($\vee_{i=1}^{n} mintermi$) are implemented simply by merging the tri-states outputs, the power consumption is likely to be a function of the number of tri-state buffers. On the other hand, the FPGA needs to be powered for the *and*/*or* gates as well as the switching systems to connect the logical blocks. Compared to a real FPGA, we have simplified our implementation, but this can only reduce power. Conversely, the FPRM is general, assuming that any Boolean Circuit can be represented as a DNF.

5. Conclusions

As part of the contribution of this work, we developed the algorithm to evaluate boolean circuits on the RM; a method to compute an optimized FPRM layout; and a method for realizing the FPRM as a tri-state circuit with comparable performance to the conventional FPGA implementation. A tri-state (MOSFET transistor) acts as a switching element in both the FPGA and FPRM. Passing a signal through a chain of k switches (that is, a chain of k source-drain connected transistors) incurs a quadratic delay of $\frac{k^2}{2} r \cdot c$ (where r is the resistance and c is the capacitance of each transistor). As a result of the reconfiguration of the FPRM, a relatively long chain of transistors can be created. Due to the short chains involved in the circuit evaluation problem discussed here, the FPRM will be able to execute the circuit evaluation process fairly quickly. In order to compare the FPRM with the FPGA/ASIC realization of $f(x_1, \ldots, x_n)$, a SPICE simulation of the FPRM can be carried out. This includes selecting the most appropriate MOSFET transistor technology to minimize signal propagation delays through the FPRM bus.

[3] The area is categorized by Units of Cells (UC), which correspond to two-input NAND gate.

This research can be furthered by comparing the synthesized results with those obtained from HLS (High Level Synthesis) of the $f(x_0, \ldots, x_{n-1})$ C-code. Analyzing other functions that can be efficiently computed using the FPRM in $O(1)$. Finally, study how the partitioning into segments affects the size of the resulting formulas, and build a decision tree that computes $f(x_0, \ldots, x_{n-1})$ on the FPRM in an even smaller size.

Appendix

A. List of acronyms and abbreviations

- **ASIC:** Application Specific Integrated Circuit
- **BC:** Boolean Circuit
- **CLU:** Configurable Logic Unit
- **CMOS:** Complementary Metal-Oxide Semiconductor
- **DNF:** Disjunction Normal Form
- **DR:** Dynamic Reconfiguration
- **DRFPGA:** Dynamic Reconfiguration FPGA
- **FDR:** Fine-Grain Dynamically Reconfigurable Architecture
- **FPRM:** Field Progammable Reconfigurable Mesh
- **HLS:** High Level Synthesis
- **LE:** Logic Elements
- **LR-Mesh:** Linear Reconfigurable Mesh
- **LUT:** Lookup Table
- **MIS:** Maximal Independence Set
- **MOSFET:** Metal-Oxide Semiconductor Field-Effect Transistor
- **NMOS:** N-type Metal-Oxide Semiconductor
- **PE:** Processing Elements
- **PR:** Partial Reconfiguration
- **PRAM:** Parallel Random Access Machine

- **RM:** Reconfigurable Mesh
- **RMBM:** Reconfigurable Multiple Bus Machine
- **RRAM:** Resistive Random Access Memory
- **SRGA:** Self Reconfigurable Gate Array
- **STCON:** st-connectivity

Author details

Esti Stein[1*†] and Yosi Ben Asher[2†]

1 Department of Computer Science, The Academic College of Tel Aviv-Yaffo, Jaffa, Israel

2 Department of Computer Science, Haifa University, Haifa, Israel

*Address all correspondence to: esterst@mta.ac.il

† These authors contributed equally.

IntechOpen

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Asher Y, B, Stein E. Evaluation of circuits on the reconfigurable mesh. In: 2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). Rio De Janeiro, Brazil: IEEE; 2019. pp. 71-74

[2] Vaidyanathan R, Trahan J. Dynamic Reconfiguration: Architectures and Algorithms. US: Springer Science & Business Media; 2004

[3] Matias Y and Schuster A. On the Power of a 2-Band Reconfigurable Network. Unpublished Manuscript. 1992

[4] Chen G, Wang B, Li H. Deriving algorithms on reconfigurable networks based on function decomposition. Theoretical Computer Science. 1993; **120**(2):215-227

[5] Nakano K, Wada K. Integer summing algorithms on reconfigurable meshes. Theoretical Computer Science. 1998;**197**: 57-77

[6] Jang J, Park H, Prasanna VK. An optimal multiplication algorithm on reconfigurable mesh. In: Proc. Symp. On Parallel and Distributed Processing. Beverly Hills, CA: IEEE; 1992. pp. 381-391

[7] Jang J, Prasanna VK. An optimal sorting algorithm on reconfigurable mesh. In: Proc. Inter. Parallel Processing Symp. Beverly Hills, CA: IEEE; 1992. pp. 130-137

[8] Elmesbahi M, KJ, Errami A, Bouattane O. Theta(1) time parallel algorithm for finding 2d convex hull on a reconfigurable mesh computer architecture. Global Journal of Computer Science and Technology. 2021;**21**:1-9

[9] Trahan JL, Subbaraman CP, Vaidyanathan R. List ranking and graph algorithms on the reconfigurable multiple machine. In: Proceedings of International Conference on Parallel Processing. NY: Syracuse University, CRC Press; 1993. pp. III–224-III–247

[10] Wang B-F, Chen G-H. Constant time algorithms for the transitive closure and some related graph problems on processor arrays with reconfigurable bus systems. IEEE Transactions on Parallel and Distributed Systems. 1990;**1**(4): 500-507

[11] Miller R, Prasanna-Kumar VK, Reisis DI, Stout QF. Image computations on reconfigurable VLSI arrays. In: Proceedings of the Conference on Vision and Pattern Recognition. Ann Arbor, MI: IEEE; 1988. pp. 925-930

[12] Trahan JL, Vaidyanathan R. Relative scalability of the reconfigurable multiple bus machine. In: Proc. Workshop Reconfigurable Arch. And Algs. Honolulu, HI: IEEE; 1996

[13] Sidhu R, Wadhwa S, Mei A, Prasanna VK. A self-reconfigurable gate array architecture. In: Field-Programmable Logic and Applications: The Roadmap to Reconfigurable Computing. Berlin, Heidelberg: Springer; 2000. pp. 106-120

[14] Hatem ME-B, Vaidyanathan R, Trahan JL, Rai S. On the communication capability of the self-reconfigurable gate array architecture. IPDPS. 2002:**500**: 0152b. IEEE

[15] Hatem ME-B, Vaidyanathan R, Trahan JL, Rai S. On designing implementable algorithms for the linear reconfigurable mesh. PDPTA. 2003: 241-246

[16] Ben-Asher Y, Stein E, Tartakovsky V. Fpga realization of the reconfigurable mesh counting algorithm. Journal of Circuits, Systems and Computers. 2021;**30**(9):2150157

[17] Giefers H, Platzner M. Armlang: A language and compiler for programming reconfigurable mesh many-cores. In: Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium. Rome, Italy: IEEE; 2009. pp. 1-8

[18] Ben-Asher Y, Stein E, Vaidyanathan R. Combining boolean gates and branching programs in one model can lead to faster circuits. In: Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2017 IEEE International. Orlando, FL: IEEE; 2017. pp. 184-191

[19] Giefers H, Platzner M. An Fpga-Based Reconfigurable Mesh Many-Core. IEEE Transactions on Computers. 2013; **63**(12):2919-2932

[20] Hauck S, Fry TW, Hosler MM, Kao JP. The chimaera reconfigurable functional unit. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2004;**12**(2):206-217

[21] Xilinx. Logicore Ip Xps Hwicap. Report DS586. San Jose, CA: Xilinx; 2010

[22] Intel Corporation. Intel® Quartus® Prime Pro Edition User Guide: Partial Reconfiguration. San Jose, CA: Intel; 2022

[23] Babu P, Parthasarathy E. Reconfigurable fpga architectures: A survey and applications. Journal of The Institution of Engineers (India): Series B. 2021;**102**(1):143-156

[24] Khan MA, Miyamoto N, Pantonial R, Kotani K, Sugawa S, Ohmi T. Improving multi-context execution speed on drfpgas. In: Solid-State Circuits Conference, 2006. ASSCC 2006. IEEE Asian. San Francisco, CA: IEEE; 2006. pp. 275-278

[25] Cong J, Xiao B. A novel fpga architecture with memristor-based reconfiguration. In: Nanoscale Architectures (NANOARCH), 2011 IEEE/ACM International Symposium. San Diego, CA: IEEE; 2011. pp. 1-8

[26] Cong J, Xiao B. Fpga-rpi: A novel fpga architecture with rram-based programmable interconnects. IEEE Trans. VLSI Syst. 2014;**22**(4):864-877

[27] Lin T-J, Zhang W, Jha NK. A fine-grain dynamically reconfigurable architecture aimed at reducing the fpga-asic gaps. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2014;**22**(12):2607-2620

[28] Ben-Asher Y, Stein E. Adaptive booth algorithm for three-integers multiplication for reconfigurable mesh. Journal of Interconnection Networks. 2016;**16**(1):1-25

[29] Rudell R, Sangiovanni-Vincentelli A. Espresso-mv: Algorithms for multiple-valued logic minimization. Proc. IEEE Custom Integrated Circuits Conf. 1985: 230-234

[30] Xilinx. The Programmable Logic Data Book. 2000. Available from: http://www.xilinx.com/index.shtml.

Edited by B. Santhosh Kumar

Data integrity is the overall accuracy, completeness, and consistency of data. Data integrity also refers to the safety of data regarding regulatory compliance, such as GDPR compliance, and security. It is maintained by a collection of processes, rules, and standards implemented during the design phase. Data governance is the process of managing the availability, usability, integrity, and security of the data in enterprise systems, based on internal data standards and policies that also control data usage. Effective data governance ensures that data is consistent and trustworthy and does not get misused. This book provides a comprehensive overview of data integrity and data governance and their myriad applications.

Published in London, UK
© 2023 IntechOpen
© gonin / iStock

IntechOpen

ISBN 978-1-83968-885-0
9 781839 688850