Universidade do Estado do Rio de Janeiro

Centro de Tecnologia e Ciências

Instituto de Matemática e Estatística

Natália Vicente Penso

Recommendations for a unified database of PWD students from a data quality perspective

Rio de Janeiro

2023

Natália Vicente Penso

Recommendations for a unified database of PWD students from a data quality perspective

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Marcelo Schots de Oliveira
Coorientadora: Prof.ª Dra. Flávia Maria Santoro

Rio de Janeiro
2023

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/BIBLIOTECA CTC/A

P418 Penso, Natália Vicente.
Recommendations for a unified database of PWD students from a data quality perspective/ Natália Vicente Penso. – 2023.
99 f.: il.

Orientador: Marcelo Schots de Oliveira
Coorientadora: Flávia Maria Santoro
Dissertação (Mestrado em Ciências Computacionais) - Universidade do Estado do Rio de Janeiro, Instituto de Matemática e Estatística.

1. Banco de dados - Teses. 2. Qualidade de dados - Teses. 3. Inclusão social - Teses. I. Oliveira, Marcelo Schots de. II. Santoro, Flávia Maria. III. Universidade do Estado do Rio de Janeiro. Instituto de Matemática e Estatística. IV. Título.

CDU 004.65

Patricia Bello Meijinhos CRB7/5217 - Bibliotecária responsável pela elaboração da ficha catalográfica

Assinatura Data

Natália Vicente Penso

Recommendations for a unified database of PWD students from a data quality perspective

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro.

Aprovada em 06 de julho de 2023.

Banca Examinadora:

Prof. Dr. Marcelo Schots de Oliveira (Orientador)
Instituto de Matemática e Estatística - UERJ

Prof.ª Dra. Flávia Maria Santoro (Coorientadora)
Instituto de Matemática e Estatística - UERJ

Prof.ª Dra. Fernanda Araujo Baião Amorim
Pontifícia Universidade Católica do Rio de Janeiro - PUC-RJ

Prof.ª Dra. Alexandra Maria Monteiro Grisolia
Faculdade de Ciências Médicas - UERJ

Rio de Janeiro
2023

DEDICATION

To my grandmother Maria de Lourdes Vicente (in memoriam).

ACKNOWLEDGMENTS

To God, for life and for always illuminating my path, especially in these last years.

To my mother Eliane, for her love, trust, advice, and for not letting me “give up” on the Master’s program in the most difficult moments.

To my sisters Patrícia, Maria Clara, and Ana Carolina, for their affection, friendship, and constant support; to my brother-in-law, Angelo, for his academic tips, support, and fans.

To my uncles Katia, Abel and Dulce, for all the support, affection, prayers and en- couragement; to my grandfather, Antônio, for the affection.

To my boyfriend Eduardo, for listening to my tireless complaints and venting about my Master’s degree, for affection and support.

To Meg, Sara and Ivy, who played a significant role in this journey, cheering me up in the most difficult moments.

To my Master’s advisor, Marcelo Schots, for the opportunities he offered me during my Master’s degree, for his guidance, and dedication.

To my Master’s co-advisor, Flávia Santoro, for the countless reviews, dedication, and orientations.

To Amanda Taylor, Maynara Govea, Maria Morais, João Esteves, Waldecir Faria, and Bruno Campos, for accepting to participate in the interviews and for their significant contributions.

To Suelen Cordeiro, for all the tips, support, motivation, and partnership throughout the Master’s program.

To professors Fernanda and Alexandra, for accepting to participate on the advisory boards and for their contribution to the research.

To my friends who always cheer for me, for all the support and encouragement.

AGRADECIMENTOS

A Deus, pela vida e por sempre iluminar meu caminho, especialmente nesses últimos anos.

À minha mãe Eliane, pelo amor, pela confiança, pelos conselhos, e por não ter me deixado “desistir” do mestrado nos momentos mais difíceis.

Às minhas irmãs Patrícia, Maria Clara e Ana Carolina, pelo carinho, pela amizade e pela torcida de sempre; ao meu cunhado, Angelo, pelas dicas acadêmicas, pelo apoio e pela torcida.

Aos meus tios Katia, Abel e Dulce, por todo apoio, pelo carinho, pelas orações e pelo incentivo; ao meu avô, Antônio, pelo carinho.

Ao meu namorado Eduardo, por ouvir minhas incansáveis reclamações e desabafos sobre o mestrado, pelo carinho e pela torcida.

À Meg, Sara e Ivy, que tiveram um papel importante nesta jornada, ajudando a me alegrar nos momentos mais difíceis.

Ao meu orientador, Marcelo Schots, pelas oportunidades que me ofereceu durante o mestrado, pelas orientações e pela dedicação.

À minha co-orientadora, Flávia Santoro, pelas inúmeras revisões, pela dedicação e pelas orientações.

À Amanda Taylor, Maynara Govea, Maria Morais, João Esteves, Waldecir Faria e Bruno Campos, por aceitarem participar das entrevistas e pelas importantes contribuições.

À Suelen Cordeiro, por todas as dicas, pelo apoio, pelo incentivo e por toda a parceria ao longo do mestrado.

Às professoras Fernanda e Alexandra, por aceitarem participar das bancas e pela contribuição à pesquisa.

Aos amigos que sempre estão na torcida, por todo apoio e pelo incentivo.

ABSTRACT

PENSO, Natália Vicente. *Recommendations for a unique database of PWD students from a data quality perspective.* 2023. 102f. Dissertação (Mestrado em Ciências Computacionais) – Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

People with Disabilities (PWD) have many difficulties and limitations that hinder their daily lives in several ways. Many of the challenges they face negatively impact stu- dent performance. Even with the advancement of medicine and the inclusion programs for PWD students, there are still a number of actions and projects to be created. These actions and projects need to facilitate and help these students keep up with regular edu- cation and develop. In this way, a unified database of PWD students would allow many analyses and projects to help with their inclusion and adaptations. This dissertation pro- poses recommendations that should be considered in order to build a unique database to register PWD students. This paper proposes using the key dimensions of data quality as a guide to recommend the core elements to be considered.

Keywords: Data quality. Data quality dimensions. PWD students. Difficulties faced by PWD students.

RESUMO

PENSO, Natália Vicente. *Recomendações para uma base de dados única de estudantes PCD sob uma perspectiva da qualidade de dados.* 2023. 102f. Dissertação (Mestrado em Ciências Computacionais) –Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

Pessoas com Deficiência (PCD) possuem muitas dificuldades e limitações que atrapalham seu dia a dia em muitos aspectos. Muitos dos desafios enfrentados acabam impactando negativamente em seu desempenho estudantil. Mesmo com o avanço da medicina e dos programas de inclusão de estudantes PCD, ainda existem muitas ações e projetos que precisam ser criados para facilitar e auxiliar esses estudantes a conseguirem acompanhar o ensino regular e a se desenvolverem. Dessa forma, uma base de dados única dos estudantes PCD permitiria muitas análises e projetos poderiam ser elaborados especificamente para auxiliar nessa inclusão e adaptações que eles precisam. Esta dissertação propõe recomendações que devem ser consideradas com vistas à construção de uma base de dados única para registar estudantes PCD. Este trabalho propõe a utilização das principais dimensões da qualidade dos dados como um guia para recomendar os elementos centrais a serem levados em consideração.

Palavras-chave: Qualidade de dados. Dimensões de qualidade de dados. Estudantes PCD. Dificuldades enfrentadas por estudantes PCD.

LIST OF TABLES

CONTENTS

INTRODUCTION

Persons with Disabilities (PWD), according to Art. 2 of Law No. 13,146/2015 (Brasil, 2015), are those who have some limitation (physical, sensory, mental or intellectual) that obstructs their effective participation in society under equal conditions (SILVA, 2022). As medicine has advanced, the ability to diagnose diseases and limitations have improved over the years. In addition, the inclusion of disabled students in schools has also gained more visibility. Due to these advances, it was possible to observe that the number of PWD students had increased almost sixfold in the last 23 years, according to the federal government's School Census (PCD, 2023).

However, it is not enough just to increase the number of vacancies for disabled chil- dren in schools. It is also necessary to ensure that the special educational needs of them are met to provide true inclusion. According to Terra & Gomes (2012), studies show how unprepared teachers feel to meet the demands of inclusion. The lack of pedagogical resources, the difficulty of knowing what disability the students have, and the pedagog- ical approach that the teachers will use with the students in class are just some of the challenges, are examples of such unpreparedness (OLIVEIRA, 2022).

Besides teachers' unpreparedness, PWD students face many barriers and challenges throughout their lives. These barriers and challenges directly affect their school perfor- mance. Some of the problems faced by these students are: (i) Difficulty in concentrating when filling out long forms or assessments; (ii) Access to digital content becomes more challenging, due to the lack of adaptation of technologies; (iii) Lack of classes and schools prepared to receive students who need adaptations, such as ramps, teacher aides, or trans- lators; (iv) Data filled out on forms by students with disabilities tends to be of low quality, since due to their limitations many answers may be entered incorrectly or left blank; (v) Need to prove his/her disability in different places and situations.

Due to these barriers and difficulties faced by PWD, some authors make recommenda- tions that could facilitate the inclusion of PWD. They also promote accessibility so that PWD can have access to all fields of interest, not only in their student life. For example, Eckstein (2022) has done research focused on the area of education in Australia, in par- ticular on the employability of students in the country's universities. His work presents 12 recommendations that should be followed in order to contribute to students' employ- ability. Neurosaber (2021), on the other hand, also focused on the area of education, in particular on multi-agent systems to recommend accessible learning objects. Some of the works that bring these recommendations will be presented in the Chapter 4.

The absence of a unified database[1] of PWD students also becomes a barrier to inclusion. The lack of such a centralized and reliable database of these students makes it difficult, for example, for teachers to plan their classes. In addition, the absence of this database impacts the lack of input into the creation of laws and public policies that invest in education for this specific audience. In addition, not having this information reliable and centralized makes it impossible to analyze data, and thus improvements cannot be proposed.

After the literature review, it was possible to observe some works intend to promote the inclusion of disabled people, especially in student life, such as in Kumar et al. (2021), Neurosaber (2021), Eckstein (2022), Tran et al. (2020). However, despite many proposals depend on the avaliable data of PWD, to the best of our knowledge, no work is concerned to the quality of databases to support addressing those issues, and no work was found in the literature that used the main dimensions of data quality to assist in this recom- mendation process. Furthermore, no work proposed building a unified database to store PWD students' data.

Based on the evidence discussed, the research problem addressed in this dissertation is: The lack to structured and reliable data reduces the opportunities of inclusion of PWD students. This will reduce some of the challenges and limitations faced by them in education. This research aims to answer the following question: What aspects are relevant for a database to support PWD?

The goal of this dissertation is to make recommendations that can be considered in the construction of a unified database to register the PWD students. Thus, this work proposes the use of the main dimensions of data quality as a guide

---

[1] Unified database is when many fragmented data sources from the same context are merged into a single central view. Unified data provides a more complete and accurate picture of data in a given context (SIGNAL, 2022).

for recommending the core items to be taken into account. It is worth mentioning that the major focus of this work are to benefit education.

This work is organized into seven chapters. Chapter 1 presents the background and motivation for the work, and the concepts related to PWD and data quality. Chapter 2 presents some data quality concepts, to complement the previous chapter, and how they relate to PWD students. Chapter 3 presents the concept of data maturity models, in addi- tion to bringing two examples of models. Chapter 4 provides the related works. Chapter 5 presents the proposal of the present work. Chapter 6 discusses the proposal evaluation. Finally, the last chapter presents the final remarks, including the contributions, limitations of the proposal and future works.

1 BACKGROUND AND MOTIVATION

1.1 **Person With Disability (PWD)**

A person with a disability (PWD) is a person with some long-term limitation of a physical, mental, intellectual, or sensory nature, which, in contact with other people, prevents them from taking part and achieving their full potential in society on an equal basis (Brasil, 2015; Assembléia Geral das Nações Unidas, 2006). More than one billion people (1 out of 7 people in the world) have some form of disability (PICHLER, 2019). The Brazilian Institute of Geography and Statistics (IBGE)'s Demographic Census conducted in 2010 shows that 23.9% of the Brazilian population – i.e., more than 45 million Brazilians – have some form of disability (PICHLER, 2019; Brasil, 2012).

Due to the significant and growing presence of PWD in society, it becomes easier to observe many limitations and difficulties faced by them in their daily lives. A person in the Autism Spectrum Disorder (ASD) commented on a professional and business identity social network that autistic people have great difficulty filling out forms (VIEIRA, 2022). During the process of applying for a job, she reported having a negative user experience using the website. The act of submitting the medical report proving her condition not only failed, but also ended up erasing all the information she had already filled in. Since the application form was not concerned with been accessible for everyone, she was not able to run for that position like the other candidates.

While the difficulties reported by this professional cause her to be ranked in a lower position in this competition, it also affects the quality of the data generated from these registration forms. Since it is difficult to fill data in a form, this directly implies that the data that they can capture has poor quality. The tendency is that part of the information collected is filled in wrongly. This is because the candidate had to enter the same information many times. This repetitive process can cause exhaustion in people with ASD, which can lead to wrong or incomplete data. Also, some information may have been left blank due to failure to submit the system or lack of attention from the candidate.

Another major challenge faced in the daily life of a person with a disability is the access to digital content. The visually impaired, for example, access websites through screen reader software and voice synthesizers, which read the web page generating audio referring to the elements that the page contains (SORRENTINO, 2013). However, many sites have a complex layout and dynamic content, making it difficult for a screen reader software to process them. Moreover, because pages are dynamically updated, screen readers are unable to identify these updates and cannot capture new information. Thus, the experience of consuming information from the internet becomes a time-consuming and complex task for the PWD (SORRENTINO, 2013).

1.2 **Classification of Diseases and Disorders**

The International Classification of Diseases (ICD), developed by the World Health Organization, is a diagnostic classification standard that defines the universe of diseases, disorders, injuries, deaths, and other related health conditions listed comprehensively and hierarchically (BIRUK et al., 2021; SANUSI et al., 2022). This classification allows for easy storage, retrieval, and analysis of health information for evidence-based decision-making, sharing, and comparison of health information across hospitals, regions, and countries (BIRUK et al., 2021). It facilitates the analysis of the general health situation of different population groups, providing information on the main issues that need to be addressed (TANNO et al., 2020).

About 70% of the world’s health expenditures are allocated using the ICD, which is used in more than 100 countries and 43 languages. As a result, changing the structure of the ICD could have significant implications for health financing and economics (TANNO et al., 2020).

The principal advantage of this classification is that it has an alphanumeric code for all existing diseases around the world. This code can be understood by any doctor, i.e., it is considered a universal language in medicine. To reflect advances in health and medical science over the years, the ICD codes have been updated several times (SANUSI et al., 2022). With each update, a new version is released and the World Health Organization adds more pathologies or changes the classification of some that were already on the list. To date, the ICD is in version 11, which gives rise to the name ICD-11.

Another important classification is the Diagnostic and Statistical Manual of Mental Disorders (DSM). DSM is a classification system for psychopathology developed in the United States by the American Psychiatric Association (APA) (MA, 2022). In order to standardize the diagnostic terms used in diagnosing mental illnesses, this classification system was developed.

The current mental health system in the United States uses the DSM-5-TR, while those in the rest of the world tend to employ a similar classification system, the ICD (MA, 2022). Besides being used by professionals in the clinical field, the DSM aims to be incorporated into other areas of work, such as legal, educational, and organizational (RESENDE et al., 2015).

Similarly to the ICD, the DSM has been updated several times to include more pathologies and make improvements to the classification. So for each update a new version was created.

1.3 **Types of disabilities**

According to the literature, in addition to the ICD and the DSM, there is a classification used to aggregate the different diseases into distinct groups, according to their characteristics (DONG; LUCAS, 2016). The first type of disease to be considered are Cognitive Disorders. Within this group, the following diseases are covered: Learning Disabilities, Attention Deficit Disorder, Attention Deficit Hyperactivity Disorder (ADHD), Dyslexia and Dyscalculia. This grouping was done because these deficiencies are related (DONG; LUCAS, 2014; DONG; LUCAS, 2016).

The other types of disease to be considered are Psychological and Physical, as defined by (DONG; LUCAS, 2014; DONG; LUCAS, 2016). In the DSM-5 (CUANI, 2021; (APA), 2014), there are more than 300 Mental Disorders cataloged. Among the many desability included in the group of psychological diseases, we highlight the most common in the literature:

- Anxiety disorders (ALVES, 2021; GANGEMI et al., 2019)
- Obsessive-compulsive disorders (GANGEMI et al., 2019)
- Depression (SCHEMITSCH; NAUTH, 2020; ARALDI-FAVASSA et al., 2005; ALVES, 2021)
- Schizophrenia (ALVES, 2021)
- Insanity (DUARTE, 2022)
- Bipolar disorder (GATTE, 2021)

The physical disease type is characterized by changes to one or more segments of the body that impair mobility, coordination, and speech, to varying degrees (Governo do Estado do Parana, 2022). According to Governo do Estado do Parana (2022), the most common physical disease are:

- Paraplegia: total loss of motor functions.
- Monoplegia: partial loss of motor functions in a single limb.
- Tetraplegia: total loss of upper and lower limb motor functions.
- Hemiplegia: total loss of motor functions in one hemisphere of the body (right or left).
- Dwarfism: is a genetic disease that causes abnormal skeletal growth, resulting in an individual whose height is much smaller than the average height of the entire population.
- Cerebral palsy: refers to a brain injury that usually occurs when the baby’s brain lacks oxygen during pregnancy, childbirth or up to two years after birth.

The classification of types of disabilities will be of great importance for ensuring the Conformity dimension of data quality, which will be discussed in the subsection 5.7.

1.4 Health data

Along with the international classifications of diseases (ICD and DSM) and the types of diseases presented in the subsections 1.2 and 1.3, there is a significant health-related concept: the Brazilian National Health Data Network (RNDS[2]).

RNDS is a Brazilian health interoperability platform created in 2020 by the SUS Informatics Department (DATASUS) (FANTONELLI et al., 2021). It aims to promote the exchange of information between the points of the healthcare network. This may allow care transition and continuity in the public and private sectors (BRASIL, 2020). Over the years, the RNDS has been constituted as a highly available, secure, and flexible information platform, in order to promote the ethical use of health data (BRASIL, 2020).

The main benefits of this platform are: (I) Improved care, with access to health information, greater transparency, care records and better provision of health services; (II) Efficiency in the management of public resources with the organization of information, action planning, increased monitoring against fraud, and the mapping of needs; (III) Innovation in health, with the use of telehealth, artificial intelligence, and appreciation of data quality (BRASIL, 2020).

The RNDS should not be understood as an information system, but as a platform to connect any information system. This includes pharmacies, laboratories, research and development centers, among others.

---

[2] Abbreviation of the Brazilian Portuguese “Rede Nacional de Dados em Saúde”

2 DATA QUALITY

2.1 **Data Quality Concepts**

Nowadays, with the great advance of technology, there has been an increase in the production and use of information (WU et al., 2018; ZHANG, 2018). The popularization of social networks and e-commerce websites, among others, contributed largely in the exponential growth of the complexity and amount of data exchanged between people and autonomous devices around the world. The need to manage this large volume of data led to the creation of novel solutions, giving rise to the concept of Big Data. Such a term describes the large volume of decentralized and heterogeneous data that are generated over time (BERMAN, 2013).

Similarly to the circumstances resulting from the Eightieth-Century Industrial Revolution, with this enormous increase in volume, the control and treatment of data are becoming more challenging each day, profoundly affecting the quality of the data.

Data quality can be defined according to different perspectives. According to Wang & Strong (1996), Pipino et al. (2002), it consists of a collection of (non-functional) quality attributes that reflect a unique perspective of data. Sidi et al. (2012), on the other hand, defines data quality as the adaptation of data to meet the user’s needs. Data quality is subjective, i.e., it depends on the use case and the domain area to which it belongs (BYABAZAIRE et al., 2020).

In the past, the data consumers were the users who generated them, who were in charge of preserving the quality of such data. However, with the arrival of the Big Data era and the wide variety of different data sources, this scenario has changed: data consumers were not necessarily the data producers, leading to a difficulty of measuring data quality (CAI; ZHU, 2015). Such difficulty becomes even greater considering that data quality is a multidimensional concept, i.e., one that must be analyzed under a holistic perspective, so as to enable its promotion and assessment in terms of established models and standards, regardless of the business rules of a particular area or domain. This multidimensional concept is difficult to evaluate, since it is hard to identify which attributes are tied to the user’s context (BYABAZAIRE et al., 2020).

There are many advantages and disadvantages related to data quality in real life. It is easy to observe some disadvantages of poor data quality in practical life (SCANNAPIECA; BATINI, 2006). For example, the incorrect delivery of correspondence is often linked to the poor quality of postal services, but when analyzed more deeply, the causes may be related to databases that have the wrong, or even incomplete, registered address.

Another example is poor-quality aerial data. When a plane is going to land on the runway, the pilot needs to know the exact moment that it will be authorized to not compromise other planes that may need the landing strip. In this way, this real-time aerial data is crucial to avoiding a plane crash, for example.

On the other hand, quality data can serve as input, for example, for the construction of draft laws. This will help the population in relation to aspects that are most needed. It also serves as input for data analysis to understand behavior and needs in each region of the country.

2.1.1 Data Quality in Organizations

Historical evidence shows that the poor quality of data has strong negative impacts on an organization. According to Redman (1998), incorrect data costs 8% to 12% of a company’s total revenue, also degrading analysis processes and any strategic decision. For service organizations, 40% to 60% of an expense can also be caused by poor data (REDMAN, 1998). In 2004, this rate of company’s total revenue increased to 20%, making the importance of data quality more evident (REDMAN, 2004).

Moreover, research by industry experts – including the Gartner Group, Price Waterhouse Coopers and The Data Warehousing Institute – also indicate the effects of poor data quality (MARSH, 2005). Some facts include (MARSH, 2005): (i) 75% of the organizations have identified costs stemming from dirty data”; (ii) “33% of organizations have delayed or canceled new IT systems because of poor data”; (iii)

“Less than 50% of companies claim to be confident in the quality of their data” – this value decreases to 15% when it comes to external data supplied to them”.

According to Chen et al. (2012), poor quality of data not only leads to wasted resources and high costs, but it can also decrease service efficiency. For example, in the context of a sales website, if the platform becomes unavailable for a few days, the data team will not be able to collect and analyze user interaction events. This can become a serious problem of service efficiency.

In addition to all the financial costs for the company, the quality of the services offered can be compromised. For example, if an online store selling football shirts has a problem updating the number of shirts available, the store may be selling more shirts than it could. This can lead to great dissatisfaction for the customer, and even a more serious legal problem.

In contrast, high data quality allows to achieve several benefits, for example, to deliver information in an agile way and proving to be a reliable organization for its customers. Correct and well-used data bring competitive advantages. An organization’s decision-making is based on information from data analysis, which provides inputs for building reliable processes (SIDI et al., 2012). Thus, quality data implies more reliable processes.

As stated by Redman (2001), organizations with the best data “can approach customers, respond more quickly (and flexibly), keep costs low, perform better, as well as make decisions faster and more confident”. However, to reach this goal, it is necessary to identify the causes of problems in advance, to address them right away.

2.1.2 Data Quality and Legal Concerns

Some non-functional attributes, handled as dimensions, are applied to better express the notion of data quality, such as integrity, completeness, consistency, precision, privacy, and security. These two latter aspects have been intensely discussed, since an increasing amount of personal and confidential data are consumed, stored, and handled by organizations, giving rise to initiatives such as the

European Union’s General Data Protection Regulation (GDPR)[3] (European Parliament and Council of the European Union, 2016) and, shortly afterwards, the Brazilian General Data Protection Law – GDPL[4] (BRASIL, 2019), first established in 2018.

This law provides for a change in the way that organizations’ operations are carried out, there are clear rules for the collection, storage, processing, and sharing of personal data (PERRONE; STRASSBURGE, 2018). Also, the customer has greater control over the use of his/her data, being able to ask the organization the access to information related to him/her, and to have the right to request the removal of such information.

Due to these region-specific regulations, organizations worldwide, such as Google, are updating their data protection terms of services to fit the new Brazilian data protection rules (GOOGLE, 2020), reinforcing the importance of handling quality aspects of the data and how they are increasingly present in people’s everyday lives.

2.2 **Data Quality and the Difficulties faced by Person With Disability**

Information collected from different sources is becoming increasingly relevant to making certain decisions and used as a guide in some situations. Thus, having quality data is a very critical premise these days. It is possible to observe a relationship between data quality and the difficulties of PWD people.

As discussed in subsection 1.1, one of the challenges faced by PWD people is access to digital content, e.g., the visually impaired people who need to use screen reader software to access the content. Comparing the quality of data, this difficulty in accessing information through websites can also result in generating low-quality data. When a user needs to send information through a website for later processing and usage, these data can be wrong or filled in an unexpected way. Occasionally, a web page may have been dynamically updated by the system when it expected to receive other specific information. However, the software that does the translation for

---

[3] In Portuguese: Regulamento Geral de Proteção de Dados da União Europeia – RGPD

[4] English translation for Lei Geral de Proteção de Dados – LGPD

PWD people does not update and does not notify the user about this change, so the user sends wrong information. This impairs the capture and quality of this data.

Valle (2012) analyzes the difficulties from the perspective of students with Attention Deficit Hyperactivity Disorder (ADHD). According to him, the school ends up excluding and harming students with ADHD because the professionals and teachers in general are not prepared. They often do not modify classroom teaching methods to meet the needs of these students. Tests that are too long and time-consuming can harm the student with ADHD, as she/he is unable to focus and perform the tests, requiring targeted support for this activity. In addition, the school environment needs well-defined rules and routines, as students with this disability are unable to understand priorities, needing to be structured, so as not to depend on memory to remember their routines on a daily basis, which would be a challenge for them.

When students have difficulty staying focused for a long time, performing extensive tests or performing other activities requiring a lot of concentration, this negatively impacts their school performance (MAIA; CONFORTIN, 2015). As a result, the quality of the data generated by this student is also impacted. Whether she(he) is answering a test or filling out a form, students with ADHD tend to struggle to concentrate in these situations for a long time and end up answering incorrectly, changing the expected answers or information, and even leaving some information blank.

According to Abreu (2012), reading and writing are skills considered fundamental in our society. From the perspective of people with dyslexia, for instance, these skills become complex. This condition often leads to considerable struggling in the production and consumption of written information.

It is difficult to keep up with the dynamism of information generation. Additionally, there is a problem with how data is made available, as well as with its quality. Information must be generated and made available in an easy-to-use and understandable format. If the information is incorrect, incomplete, duplicated, or out of date, it is of no use, since it is poor quality information.

2.3 **Data Quality in Education**

The Brazilian National High School Exam (ENEM)[5] was created in 1998 to evaluate students' performance after completing basic education (SILVEIRA et al., 2015). After a while, many Higher Education Institutions[6] began to use it as a selection tool for students to get into universities (JUNQUEIRA et al., 2017). In 2009, ENEM took on another dimension with the implementation of the Unified Selection System[7] (VIGGIANO; MATTOS, 2019). There was a reformulation in the structure of its test, in addition to initiating a greater inclusion of people with disabilities, thanks to the public policy of specialized care (DUARTE, 2020). Many works in the literature present different perspectives when analyzing the ENEM database.

The ENEM database is an open database and available at the official website for consultation. Among the information available in this database are: the year of the test; information related to the student's region of residence such as state and municipality; place of birth; marital status; race; type of school he studied (public or private); if he has any disability; what disability he has fits (blindness, low vision, deafness, physical disability, mental disability, autism, ...); if you need any specialized assistance to carry out the exam (special chair, left-handed chair, furniture for the obese, braille, special room, specific material,...); and grades by subject.

An important study proposed in Duarte (2020) aims to analyze the performance of students with disabilities and evaluate the effect of the public policy of specialized care on their performance, in ENEM 2019. The relevance of this research is to understand the performance characteristics of students with and without disabilities in a basic education assessment exam, obtaining subsidies to propose new public policies that help to improve the educational performance of students with disabilities and, at the same time, greater social inclusion. In addition, the work intends to to assess the impact of the public policy of specialized care on the performance of students with disabilities, making it clear whether there has been an improvement in performance, or if this public policy could undergo some changes to contribute more efficiently in this scenario.

The study made by (DUARTE, 2020), proved that students with disabilities, who received a public policy of specialized care, had better average grades than the

---

[5] ENEM is the acronym of the definition in Portuguese: Exame Nacional do Ensino Médio.
[6] In Portuguese: Instituições de Ensino Superior – IESs.
[7] English translation for Sistema de Seleção Unificada – Sisu.

average grades of students without disabilities who did not receive this public policy. So, the need for education is reinforced, both to increase the inclusion of people with education, and to increase investment in education, as in other public policy.

Several other works use the ENEM database for various analyses, not only for students with disabilities, but for all students in general to understand factors such as parents' education, family income is more than two minimum wages, and if the student has access to a computer at home, they directly influence their performance.

In Viggiano & Mattos (2019), the main focus of the analysis is to understand the performance of students in ENEM in different Brazilian regions. This research showed that, with the exception of Writing, all areas of knowledge had a considerably higher performance in the South and Southeast regions, compared to the North and Northeast. Therefore, the region influences the performance in the ENEM. Based on that, the study indicates that despite the implementation of diagnostic evaluations by the government, the creation of ENEM has brought limited results, showing the need to expand policies to improve Brazilian education, both in public and private schools (VIGGIANO; MATTOS, 2019).

In Noguera et al. (2020), the ENEM database is used to understand the performance of women in the ENEM, considering the areas of knowledge of this exam related to the exact sciences. Its main objective is to assist educational decision-making in proposing specific actions in the field of education, with a focus on gender balance in the exact sciences. Based on the analyzes carried out, it is possible to notice that the performance difference between female and male participants is not so great when compared to the imbalance of these sexes in the courses in the area of exact sciences and in the professionals who work in these areas. So, the study raises the need for the possible creation of projects or even public policies to assist in this gender equality.

Based on those examples, it is possible to perceive that there are different possibilities of analysis in the ENEM database. Decision making, creation or maintenance of public policies and even the creation of bills can be made based on these data. So, as these data is of great importance for various actions to improve education, it is necessary to guarantee its quality, i.e., data should be reliable so that no decision is taken in the wrong way. However, some studies also shown that ENEM database has not the expected quality.

Studies such as Alcantara et al. (2015) show that in Brazil, there are problems in relation to educational data due to the lack of completeness, quality of information and inadequate data formats for reuse. These problems create barriers to the development of new solutions and even government initiatives. Bandeira et al. (2015) state that the great problem of the quality of educational data makes the consumption of this data to be done through a lot of effort and cost.

After the literature review, it is possible to note the importance of having a unique database of PWD students. Many studies in the literature used ENEM data for analyses. But what about students who have not taken ENEM? What about children who are not yet old enough to participate in this selection process? It is critical that data analyses are conducted early in the lives of PWD students. This should not restricted only to the socio-economic information provided by ENEM. PWD students face challenges and limitations at an early age.

3 DATA-ORIENTED MATURITY MODELS

As seen in previous chapters, data quality is a very critical and embracing area. As a result, data quality models were created to assist in its management. In the following sections, two main data quality models are discussed: the Data Management Capability Model and the Data Management Maturity Model. Additionally, the main dimensions of data quality will be presented. At the end, the relationship between PWD challenges and data quality are explored.

3.1 **Data Management Capability Model (DCAM)**

The Data Management Capability Model (DCAM) was developed by the Enterprise Data Management Council in 2014. DCAM is a data management capability model that combines the best practices related to the management of data content. Also, this model determines the scope needed to establish, enable, and maintain mature data management (COUNCIL, 2014).

Since data are generated in different places and at all times within an organization, there is a need for strong data management. Given this, the DCAM model assists in the identification of a data control structure and defines what resources are made available so that data management becomes a critical part of an organization’s day-to-day operations. The DCAM is organized into eight core components: Data Management Strategy, Data Management Business Case and Funding Model, Data Management Program, Data Governance, Data Architecture, Technology Architecture, Data Quality and Data Operations (COUNCIL, 2014).

The Data Management Strategy addresses for an effective data strategy, its importance in the organization, and how an organization needs to self prepare to be able to implement it. The Data Management Business Case and Funding Model broach the creation of the business case and reinforce the importance of involving key stakeholders for approval. The Data Management Program addresses which are essential for a data management program to be sustainable. Data Governance is responsible for determining an operational model to be followed and emphasizing the importance of the procedures and policies that must be used to achieve alignment between the selected parties (COUNCIL, 2014).

Data Architecture is concerned with how data will be defined and related. The Technology Architecture is responsible for the link between the data and the physical IT infrastructure needed to support it. Data quality is related to the idea of the data being adequate to the standards they were related to, and to the data control and manage- ment processes. Data operations are related to the data life cycle, i.e., they assist in the definition of data life cycle processes, in addition to assisting in the integration of data management with the organization’s day-to-day activities (COUNCIL, 2014).

This model requires a controlled environment to be created in order to assist in data management. To create this environment, the organization’s executives, together with the data owners and Corporate Audit, establish policies and standards that must be respected.

3.1.1 Data Quality Program

The Data Quality Program is one of the eight main components of DCAM. This program is responsible for defining the goals and action plans that ensure that the data is of sufficient quality to assist in the organization’s strategic objectives, i.e., create a data quality strategy. And it is through this program that DCAM creates a shared culture of data quality that will allow the integration of all areas of the organization with the data quality strategy.

The crucial aspect of the program that must be emphasized is that it is not a process in itself, but it is the union of all the processes in the data availability chain that is responsible for quality assurance, and for ensuring that the data will meet as needed by customers.

Besides, this program is developed in line with business objectives and measured based on the data quality dimensions that were previously defined (COUNCIL, 2014). The prioritization and identification of these main dimensions of data quality are a very important pre-requisites for good data quality management.

There are many dimensions of data quality that can be analyzed, but in DCAM the most emphasized dimensions are completeness, coverage, conformity, consistency, accuracy, duplication, and timeliness.

This program also proposes that a profile of the current state of data quality be created regularly. The central idea is to assess what the data standards are, to check if there are anomalies and similarities between the new data and those that were already stored in the database (COUNCIL, 2014). With this, it is possible to analyze whether the new data are in line with the business requirements and the data quality requirements that have been established. Thus, based on this analysis, improvements and adjustments can be suggested so that the new data undergoes a cleaning process that will leave them as expected.

This data quality program identifies the data elements that are most required by the organization, helping to prioritize and guarantee their quality. Besides, the data quality process must be documented, operated and validated frequently, through quality data management, and monitored by an audit process (COUNCIL, 2014).

3.2 Data Management Maturity Model (DMM Model)

The Data Management Maturity Model (DMM Model) was inspired on the Capability Maturity Model Integration (CMMI) of the CMMI Institute, in 2014. The DMM Model gathers the best practices to establish, construct, monitore and optimize data management across the data life cycle (CMMI Institute, 2014). Besides, this model also offers an extensive framework to assist in data quality management.

This model can be applied to any organization and any data management objective. DMM is divided into six categories: Strategy, Governance, Data Quality, Operations, Platform and Architecture, and a Supporting Process. And this categories are divided into 25 process areas: 20 data management process areas and 5 supporting process areas.

Another very interesting aspect of this model is that it is a flexible model, allowing the specific need of each organization to be met with it. This model flexibility exists because it is possible to implement any combination of process areas or categories, i.e., it is possible to evaluate capacity or maturity in a single process area, a set of process areas, a category, a set of categories or the entire model (CMMI Institute, 2014).

DMM divides the maturity capacity of data quality into 5 cumulative levels (CMMI Institute, 2014). At Level 1 (Performed), data quality practices are done in a restricted way, often by an area or person, and in general without much control. In Level 2 (Managed), practices have already started to be carried out with a certain level of management, with processes to be followed. Going up to Level 3 (Defined), the practices are standardized and well-defined through a prescriptive process, presenting greater control of management and governance. At Level 4 (Measured), the organization uses quantitative techniques to measure its processes and manage its practices. When the last Level (Optimized) is reached, the organization is able to focus on performing only the optimization of its practices, in a continuous improvement.

Capacity and maturity assessment is done differently. For a process area to be rated at Capacity Level 4, for example, it must perform all practices provided for Levels 1, 2, 3 and 4, mandatorily. As for the maturity assessment, it requires that all process areas that are being measured in a given category be performed at the function’s capacity level and also at the same level in infrastructure support practices. That is, for a process area to be rated at Maturity Level 4 in a specific category, all appropriate practices must be performed for Levels 1, 2, 3 and 4, and all infrastructure support practices must also be implemented performed at Levels 1, 2, 3 and 4 (CMMI Institute, 2014).

3.2.1 Data Quality in the Context of DMM

Data Quality is one of the six main categories of DMM, and is responsible for the set of best practices for detecting, evaluating and cleaning data to ensure that the data meets quality requirements and business requirements. This category is divided into four process areas: Data Quality Strategy, Data Profiling, Data Cleansing and Data Quality Assessment.

The Data Quality Strategy is the process area responsible for defining goals, plans and objectives to improve data quality. Also, it is responsible for decreasing the amount of information that is redundant, obsolete and trivial in the data (CMMI Institute, 2014).

According to the DMM model, a good data quality strategy is one that is based on the main existing quality problems and is aligned with the strategic objectives of the business. Besides, it is very important that the entire organization is aligned with the data quality strategy, often requiring a cultural change in the organization (CMMI Institute, 2014).

In developing the data quality strategy, it is very important that the main dimensions of data quality are prioritized and identified, just like in the DCAM model. In the DMM model, the main dimensions of data quality that are emphasized are: Accuracy, Completeness, Coverage, Conformity, Consistency, Duplication, Integrity, and Timeliness.

The Data Profiling is often done just like in the DCAM model. In this process area, the new data are analyzed and compared to the data already existing in the databases or compared to the metadata and documentation of how the data is expected, so that any anomaly or difference can be detected and treated.

The Data Cleansing is the process area responsible for correcting data that is out- side of standards or that does not comply with data quality requirements and business requirements. For this, the dimensions of quality that were previously defined and the business rules assist in this task.

The Data Quality Assessment is responsible for assessing data quality based on pre-defined quality rules and concerning business requirements. The assessment begins to be carried out by the most critical and important data of the organization, such as financial data. Besides, goals, limits, and metrics are established to facilitate the evaluation and dissemination of the results obtained.

3.3 **Dimensions of Data Quality Models**

The term dimension has different meanings depending on the context that is used. In this work, dimension is the set of characteristics or objectives that are used to assist in the assessment of data quality.

Table 1 presents how the DCAM (COUNCIL, 2014) and the DMM Model (CMMI Institute, 2014) define the dimensions of Data Quality.

Table 1 - Dimensions of Data Quality Models

Dimension	**Definition**
Accuracy	Helps to measure whether the analyzed data matches or are close to the actual phenomena they represent.
Completeness	Concerns the availability of the needed data attributes.
Coverage	Addresses the scope of the necessary data records.
Conformity	Ensures that the data meets the required standards.
Consistency	Refers to the compliance with the required standards and rules.
Duplication	Relates to the redundancy of records or attributes.
Timeliness	Assesses whether certain data is available when neces-sary.
Integrity (only in DMM)	Checks the accuracy based on the data relationships.

Source: The author, 2023.

Table 2 presents the definitions of data quality dimensions used throughout this work.

Table 2 - Definition of the Dimensions of Data Quality in the literature

Dimension	**Definition**
Accuracy	Verifies that the modeled data is in line with the real world, i.e. it verifies that the information is accurate enough and close enough to reality (EPPLER, 2009; JAYAWARDENE *et al.*, 2015; SMARTBRIDGE, 2020). Degree to which the real-life situation is different from its representation (NICHOLSON *et al.*, 2021).

Completeness	Indicates the degree to which values are present in a data collection (SIDI *et al.*, 2012; BATINI *et al.*, 2009; REDMAN; GODFREY, 1997). Mainly related to the existence of empty values (SERHANI *et al.*, 2016).
Coverage	Indicates which real-world entities the data represent (FILL *et al.*, 2022). A measure of data availability and comprehensiveness compared to the total data universe or population of interest (SIDI *et al.*, 2012).
Conformity	Determines how well data conforms to a specified format (JAYAWARDENE *et al.*, 2015; RANJIT; KAWALJEET, 2010). Checks that the data is following the set of standard data definitions such as data type, size, and format (SMARTBRIDGE, 2020).
Consistency	The data in all systems should reflect the same information, synchronized with each other across the enterprise (SMARTBRIDGE, 2020). Checks that the data is presented in the same format and compatiblewith the previous data (SIDI *et al.*, 2012; WANG; STRONG, 1996).
Duplication	Provides a measure of unwanted duplication for a particular field, record, or data set (SIDI *et al.*, 2012; MCGILVRAY, 2008).
Timeliness	Refers whether the information is available when it is expected and needed. (SMARTBRIDGE, 2020). Refers only to the delay between a change of a real world state and the resulting modification of the information system state (SIDI *et al.*, 2012; SCANNAPIECA; BATINI, 2006).
Integrity	Indicates the ability to link related records with each other, i.e., it is responsible for verifying the link between the data (RANJIT; KAWALJEET, 2010).

Source: The author, 2023.

3.4 Limitations and challenges of PWD people related the dimensions of data quality

The previous subsections discussed the importance of data quality models and data quality dimensions for ensuring and managing data quality in general. The purpose of this subsection is to examine how the challenges and limitations of PWD relate to the dimensions of data quality discussed in the chapter.

People with disabilities have some limitations and challenges in their daily lives that require special attention, as discussed in sections 1.1 and 2.2. For example, one of these difficulties was related to people with autism who have difficulties in filling out forms, and with the lack of accessibility that is often found in this activity, they end up not being able to participate in certain activities that require this filling. Considering this difficulty, and that the data quality dimension of **Consistency** has the role of ensuring that the data in all systems reflect the same information and are synchronized with each other everywhere they are stored (SMARTBRIDGE, 2020), this dimension should be considered to alleviate this problem.

From the perspective of education and how to facilitate filling in forms, it would be interesting if the student had a single record of his data and illnesses. Thus, much of the information that needs to be filled out about their disabilities would already be registered and would be automatically identified when the student entered unique personal data, such as the SSN (Social Security Number)[8]. Therefore, the amount of information that the person would need to fill in would be reduced, easing the effort to complete this filling.

If a student has a permanent illness that was diagnosed in high school, for example, that information should be associated with her-him when he enters college. So all educational platforms should have consistency between them, and if the student needed to update one of them, they should all update automatically. If a student needs to transfer from one college to another, that important information about her-his illness should al- ready be one of the first to be taken into account, as it is often necessary to adapt the physical space of the university to give that student accessibility. Or suddenly include experienced professionals to help them in the classroom.

Another proactive issue that the **Consistency** dimension could contribute is to facilitate the analysis of data from PWD students, because if all the necessary information had in one place and with quality, many actions, bills, and other initiatives could be taken to help improve accessibility and follow up on these actions. In addition, if the databases of the educational institutions do not speak to each other, i.e., they do not have a relationship between them, the data are dispersed and the student would have to remember and provide all the documentation again to prove certain characteristics and conditions, and this is unnecessary bureaucracy as this

---

[8] English translation for Individual registration – CPF

has already been proven once. So if this process were somehow unified, consistency would be guaranteed and would already help people with disabilities a lot.

In addition, if a PWD student is supposed to participate in an academic event, the place where the event would be hosted could already proactively analyze the database of students enrolled in the event and provide the necessary items for accessibility, instead of always asking in the questionnaire if that student would need some help. This would help make accessibility more transparent. Furthermore, students often do not feel comfortable talking about their illnesses in certain ways, even to avoid prejudice.

Considering also this problem of the difficulty of filling in forms, and adding to other problems presented in the sections 1.1 and 2.2, both in relation to the difficulty that students with Autism Spectrum Disorder (ASD) have in taking long tests, and the barriers faced by students with dyslexia in competences simple as reading and writing, the data quality dimension Completeness, responsible for verifying the availability of the necessary data attributes (CMMI Institute, 2014), should be considered to alleviate this problem.

Students with ASD in addition to having a lack of attention due to their illness, they avoid activities that demand a great mental effort. So, when filling out a registration form or taking a test, in addition to filling in wrong data due to lack of attention and concentration, they can leave incomplete information because they can’t stay long in the same activity. Then, relating the data quality dimension **Completeness** with the data quality dimension **Integrity**, if there was a single record about the student, as seen above, he would have less information to fill in the form, reducing the problem of completeness.

Students with Dyslexia have difficulty understanding the written text, and depending on how an application form is prepared or the test questions, these students could often leave information blank because they do not understand what should be filled out in each case. In addition, students with dyscalculia do not understand the meaning of numbers, so if enrollment forms, tests or any other way of capturing data have questions involving numbers (examples: number of people who live with them, average family income, number of household items,...), they will have difficulty and possibly leave information unfilled.

In this way, the **Completeness** data quality dimension could help to guarantee that the data is filled, in addition to guaranteeing their validation. For example, in competitive exam answer cards, a way to ensure that a given question answer left blank, student wanted to leave it blank, was not forgotten. Another issue that can be analyzed is about students who have some visual impairment that hinders the marking of answers in the "small squares" model, as is done in most contests. This marking should be made in a more accessible way so that the student does not necessarily have to involve another person in this process, as the idea is to try to make the person as independent as possible so that they also develop autonomy and confidence.

Regarding the format of the tests is to consider the issue of time spent, the number of questions that must be performed for each day of application of the test, and if there is any way to provide an environment in case the person feels any discomfort during the completion, and that she can have the support she needs so she can resume. People often tend to say that people with disabilities cannot afford to go to a certain education because they are limited by their illness, but in fact, what needs to be done is to adapt current education systems so that they can be accessible to all.

Returning to the idea of having a single record about the personal data and illnesses of each student, the **Timeliness** data quality dimension, that verifies that data be available when needed, could also be considered to help with this issue. Data referring to PWD students can be temporary or permanent. Temporal would be cases in which the person’s limitation was for a determined time, for example, if he had an accident and is temporarily using a wheelchair. Permanent refers to diseases or conditions that have no cure, such as Multiple Sclerosis, which is a disease that still has no cure, just treatment. Thus, the automatic updating of information in an integrated manner also contributes to these aspects.

The **Duplication** data quality dimension, which checks whether attributes or records are in redundancy, could also contribute to some of the problems faced by PWD people, as seen in sections 1.1 and 2.2. This dimension can include the issue of versioned data, i.e, if a student with a cognitive issue forgets that he has already submitted a certain form, he may end up submitting the form again and even with divergent information and generating duplicate or inconsistent data. In addition, if a student had not yet been diagnosed with a disease at the time of filling out the application form, but was diagnosed after that, there should be a way to update this information without duplicating such data.

Furthermore, the **Coverage** data quality dimension, related to the availability and comprehensiveness of data compared to the total universe of data or population of interest (SIDI et al., 2012), could also contribute to some of these problems faced by people with disabilities. It is important that education-related public tender forms are able to cover all applicants, including PWD students. Since entries in public tenders often do not consider accessibility for people with disabilities, either through forms that are difficult to interpret, or even websites without accessibility for some disabilities, coverage of the entire population is compromised. It might cause segregation and even discrimination of a part of the population.

Another dimension of data quality that can also be considered to help alleviate some of the problems that people with disabilities experience is **Conformity**. As this dimension is responsible for ensuring that the data meets the required standards (CMMI Institute, 2014), i.e., it determines to what extent the data conforms to a specified format (JAYAWARDENE et al., 2015), it could contribute with regard to the difficulties in processing information. Students with cognitive difficulties, such as ASD, dyslexia and dyscalculia, when they need to fill out forms with open fields that need to follow a certain pattern, may not be able to follow certain patterns and end up filling in the data without conformity as they should, impairing the quality of the data to be processed and analyzed.

Moreover, using open fields in forms for people with cognitive difficulties may not be the best option to capture this information. For example, there are many websites that have a checker for the zip code field to see if the person has filled in the value as expected.

However, it would be more interesting and make filling in this information more accessible if for each digit of the zip code there was a box or a separate marking to make it clear where each digit should be filled. This would minimize forgetfulness or confusion in filling in such numbers. In this way, making it more explicit what and how each field should be filled in could help people with disabilities and even facilitate such filling.

Another example that this concept fits is about answers to test questions. If the question has more than one answer linked to it, and the field for the answer is open, perhaps people with disabilities may be confused, because they may not remember that they had to answer each item in the question, and end up being harmed in the test.

The Accuracy data quality dimension verifies that the modeled data is following the real world, i.e., it demonstrates that the information is accurate enough and close enough to reality (EPPLER, 2009; JAYAWARDENE et al., 2015). Thus, making the connection with the problems listed in sections 1.1 and 2.2, the accuracy could contribute in some ways, one of which is about the filling of data by third parties. Some people have disabilities or limitations that prevent them from filling out forms or any type of application by themselves. But would these data filled in by third parties have the same quality if they could be filled in directly by students? Often third parties end up making mistakes or not fully understanding the real needs of the student and may miss some critical information. Thus, this data must be validated in some way, for example, by contacting the person who has the disability.

Accuracy could also help in validating with a reliable source that a particular person has a particular disability. Any type of registration that the person has to make for some evaluation, to manifest that she is a person with a disability, she has to prove that she is a person with a disability. Thus, there should be some mechanism that allows the automatic verification of this information. With accurate data guaranteed, an API could be created to integrate with any education-related online application. Thus, the person would only need to put a code or a validation QR code of some document that proves their disability and this information would be validated to a reliable database, ensuring proof of such disability. So, this facilitates the registration process, facilitates the guarantee that the person has a disability, and the process would be less bureaucratic and less limiting.

Considering all these evidences of the need to improve the situation of people with disabilities, this dissertation proposes recommendations for the creation of a unified database designed for students with disabilities.

4 RELATED WORK

The following subsections describe the works related to this dissertation. These works make recommendations to improve the lives of people with disabilities, as does the present proposal. Some authors, such as Kumar et al. (2021), Neurosaber (2021), Eckstein (2022), Tran et al. (2020), make recommendations that could contribute to facilitating the inclu- sion of people with disabilities. In addition to inclusion, these recommendations also aim to facilitate and promote accessibility so that PWD can have access to all fields of interest.

4.1 Recommendation for Improving Disability Employment

Several limitations hinder the daily lives of people with disabilities. The reason for this is that most of the world we live in was designed by people without disabilities. Despite their lack of knowledge, these people fail to recognize the challenges faced by people with disabilities. Whether in medical, educational, or even professional areas, these people face different barriers. According to Tran et al. (2020), sustainable employment of people with disabilities is an example of these challenges. In Australia, for example, around 53,4% of disabled people are employed, in contrast to 84,1% of non-disabled people. It demonstrates the difficulty of integrating disabled people into the workforce.

To help reduce these barriers in the professional area, Tran et al. (2020) proposes recommendations that increase the employability of job seekers with disabilities. Also, according to the author, most recommendation systems implemented for the employment service use traditional recommendation models that use models where recommendations are optimized based on results. If a particular position requires a driver’s license as a differential. The candidate does not have one, he recommend that the candidate obtain a driver’s license.

Tran et al. (2020), on the other hand, make recommendations based on improving the probability of employment. Suppose, for example, there is a vacancy and a candidate with a license has the same chance of getting the job. This is regardless of whether or not they have a license. Thus, it is not recommended that the candidate obtain a license, as there would be no improvement in employment prospects.

Thus, through the methodology of causality, i.e., which has the use of the causal effect, Tran et al. (2020) estimate the improvement in the employability of people with disabilities through recommendations that can improve the probability of getting such a job. In addition, recommending people with disabilities for certain positions requires making critical decisions that can be done through this methodology.

4.2 Recommendation for inclusion of university PWD students in the labor market

All people have the right to work. However, some continue to face barriers to getting a job because they have a disability, especially Australians (ECKSTEIN, 2022). In addition, university students with disabilities also have a disadvantage on the labor Market compared to others. Although universities are not as influential on the labor market, some actions can be taken to help PWD students make a better transition to the labor market and secure employment that is relevant to their studies (ECKSTEIN, 2022).

Given this disadvantage of university students with disabilities in the labor market, Eckstein (2022) recommends some ways to improve the provision of university career services for students with disabilities in Australia. To create these recommendations, data on students with disabilities, staff, and other academic roles were collected.

According to the data collected, there are universities in Australia that already provide targeted career support to PWD students. However, they represent only 55,8% of universities in Australia, demonstrating significant opportunities for improving inclusion in the country’s education. Eckstein (2022) indicates that with systemic changes within universities, it will be possible to make the employment scenario more inclusive for PWD students, valuing their aspirations and skills.

Thus, to achieve this systemic change, Eckstein (2022) makes 12 recommendations that must be followed. Listed below are some of the main recommendations:

- The education sector, in collaboration with other institutions, investigates the provision of a national career strategy for PWD students to guide specialist services.
- Specialized disability bodies promote the sharing of information on employability and the challenges faced by PWD students on the labor market.
- Universities use their contacts with national professional associations to develop collaborative partnerships with employers. And, in this way, they can promote trust and mutual understanding of the benefits of students with disabilities.
- Universities collaborate to investigate ways to provide cooperative support for the career development of PWD students.
- Universities partner with Disabled Employment Services (DES) providers for the benefit of the university sector.
- The government should review the current funding model for DES providers to improve compensation for their investment in supporting college students with disabilities.

4.3 Recommendation for Accessible Learning Objects

School inclusion is responsible for promoting diversity and ensuring access to education for all children, whether they have a disability or not. Promotion of inclusion is not an easy task, as it must overcome many obstacles. The main agents of this transformation are the teachers, who also deal with these challenges daily in the classroom (NEUROSABER, 2021). This happens because they receive students with different levels of development, disorders, and diseases. So, finding the best teaching strategies and methodologies that are suitable for all students becomes a very challenging task.

This challenge motivated Mourão & Netto (2019) to develop a Multi-Agent Recommender System for recommending accessible learning objects, designed for students with disabilities and professors in the computing field. To develop this system, questionnaires answered by teachers in the area were used. With the answers to these questionnaires, qualitative analyses were carried out to classify, organize and make available the most appropriate Learning Objects according to the teachers' preferences.

Using the information provided by the teachers, Mourão & Netto (2019) propose the best content recommendations based on the type of disability (visual, physical, auditory, or cognitive), the discipline, and the content to be taught, based on the best recommendations that the agent can make.

The recommendation system provides a new motivational practice through accessible educational content, which significantly contributes to the teaching-learning process of the computing course and inclusive education (MOURÃO; NETTO, 2019).

4.4 Recommendation to improve the medical rehabilitation of PWD

According to Kumar et al. (2021), many people with disabilities believe that their situation negatively affects their chances of living normally like other people, such as going to school, working, and enjoying family life. The quality of life, in addition to being compromised for people with disabilities, is also for their families. Because the social segregation of PWD is also generalized, bringing negative consequences of social stigma that affects the whole family.

Given these challenges faced by people with disabilities, Kumar et al. (2021) performed a survey to study the sociodemographic profile of people with disabilities who were admitted to a rehabilitation hospital in India. This study aimed to understand the barriers faced by PWD regarding the health services and rehabilitation available to them, and thereby recommend actions that could be taken to improve their living conditions during rehabilitation.

After performing qualitative analyzes during the research, Kumar et al. (2021) suggested 7 recommendations that should be followed:

- Facilitate the integration of systems and services for PWD through advocacy-related services. As an example, norms, policies, and strategies.
- Invest in specific programs and services for PWD, as in addition to traditional services, some people with disabilities may need more specific services for each type of disability. It is recommended that this investment be carried out jointly with people with disabilities. Consequently, because they have experience in the practice of their challenges, they can provide information about their problems and suggest possible solutions.
- Train health care providers and program managers so that they can be better prepared to care for PWD.
- Educate children with disabilities as close to the education that children without disabilities receive.
- Increase public awareness and understanding of disability through social marketing campaigns.
- Improve the generation of PWD-related data, as they will help in the planning and execution of appropriate measures to address the problems of people living with disabilities.
- Support and encourage research on disability.

4.5 Comparative Analysis

The literature includes many studies related to people with disabilities. As presented in the previous subsections, some of them suggest recommendations to reduce barriers facing PWD and to help promote accessibility for all.

As there are many barriers and different areas that need to be studied to promote accessibility for people with disabilities, each author chose a problem to be worked on. This was within a specific area. Kumar et al. (2021) did their research focused on the medical field, in particular on the rehabilitation of people with disabilities. In his study, Neurosaber (2021), focused on the area of education, particularly multi-agent systems to recommend accessible learning objects. Eckstein (2022) has done research focused on the area of education in Australia, in particular on the employability of students in universities in the country. Finally, Tran et al. (2020) focused its studies on the professional field, specifically to assist people with disabilities to find work.

To be able to provide the appropriate recommendations, a quantitative analysis of different databases of people with disabilities was conducted in Kumar et al. (2021), Eckstein (2022). Kumar et al. (2021) focused on databases that had the sociodemographic profile of people with disabilities who were admitted to a rehabilitation hospital in India. (ECKSTEIN, 2022) focused on data about students with disabilities, staff, and other academic functions.

On the other hand, in order to be able to propose the necessary recommendations, Neurosaber (2021) carried out qualitative analyses based on questionnaires developed by him and applied to education teachers. In Tran et al. (2020), his recommendations have been based on studies of the literature.

As can be seen, several authors have worked on research to improve the living conditions of people with disabilities in various spheres and perspectives. However, none of these works considered the "digital" sphere, and more particularly, the sphere of data for these people. In addition, no work was found in the literature focused on the construction of a unified database for students with disabilities. Therefore, the present proposal is justified.

The main contribution of this work is in the area of data quality, with the main focus being on the area of education in the first instance. However, this work can be useful in other areas, such as the medical field, for example.

5 PROPOSED APPROACH

People with disabilities have some limitations and challenges in their daily lives that require special attention, as discussed in Section 2.2. In addition, as presented in Section 2.3, many analyses can be carried out in the education field, mainly aimed at students with disabilities, bringing other limiting factors and challenges faced by them.

As a way of contributing to mitigate these problems, this dissertartion argues that it would be interesting to have a central database to identify students with disabilities and those who have some type of limitation that hinders their development in an education environment. The Brazilian government uses a tool called CadÚnico, a tool that allows to record and manage families living in poverty and extreme poverty in the country (MUNIZ; BATISTA, 2022).Similarly, a unique database of students could be used to assist in (i) government actions for PWD students, (ii) creation of inclusive lesson plans and activities, (iii) assist educational and health professionals in developing strategies to empower PWD students in performing education activities according to their limitations, as well as (iv) providing relevant information to institutions so they can adjust themselves based on the diversity of PWD students, among other potential benefits.

To set up this database, it is necessary to consider several aspects, not only in terms of its structure, but also to its content from the data quality dimensions viewpoint, thus ensuring its effectiveness.

Thus, the purpose of this work is to make recommendations that can be considered in the construction of a unified database to register the PWD students. This work proposes the use of the main dimensions of data quality as a guide for recommending the core items to be taken into account.

Based on these recommendations, a database would be able to achieve the following goals:

- Serve as a central, trusted data source for the country’s students.
- Assist in data analysis aimed at PWD students or who have some type of limitation, making them more assertive.

- Assist in measuring the results of public policies aimed at PWD students.
- Assist in the monitoring of investments in the area of education specific to this audience.
- Be a facilitator when registering for exams, colleges, faculties, and any other variety educational institution.

It is worth reinforcing that the scope of this work does not include the construction of the database, but solely to provide the recommendations to assist in its construction. The actual development of such database and an effective evaluation of its benefits depends of community effort, involving additional academic efforts and/or the contributions of practitioners from government and industry.

5.1 The Brazilian Inclusion Registry

In July 2015, the Brazilian Law for the Inclusion of Persons with Disabilities[9] was enacted, aimed at ensuring and promoting, on equal terms, the exercise of fundamental rights and freedoms by PWD, aiming at their social inclusion and citizenship (Brasil, 2015). This statute established the National Registry of Persons with Disabilities[10] initiative. It consists of an electronic public record with the purpose of collecting, processing, systematizing, and disseminating geo-referenced information to allow the identification and socioeconomic characterization of the PWD, as well as barriers that prevent the realization of their rights (Brasil, 2015).

The expected outcome is to contribute to the planning and development of public policies for PWD, to promote these people’s access to their rights, in addition to being a source of data for studies and research (Brasil, 2022a).

An interview conducted by Santos (2018) interpreted the Inclusion Registry, at first sight, as a registry of people with disabilities. However, it will not be a

---

[9] Statute of Persons with Disabilities
[10] Inclusion Registry

registration in itself, but a crossing of existing databases, which requires data from other agencies, ministries, federal government agencies, etc.

This subject matter has matured and been improved over the years; in March 2022, was launched the first stage of the inclusion registry (Brasil, 2022b). This step put forward the Certificate of Persons with Disabilities, a public certificate that citizens with Disabilities can download to certify their condition in front of the public and private bodies and entities. However, not all disabled people were awarded this certificate, but only those who (i) receive a financial support called Continued Benefit – towards those who live with lower income – or (ii) those who receive the Retirement of Persons with Disabilities (Brasil, 2022b).

The second stage of this inclusion registration was scheduled to December 2022, in order to allow every people with disabilities to apply for the certificate, subject to undergoing a new biopsychosocial assessment (Brasil, 2022b).

Since this registry's database is still under development, no literature is available, no data project has been developed, and no metadata has been developed.

In the following subsections, recommendations will be presented in the context of each data quality dimension. In addition, for each dimension a summary table of recommendations will be presented with the following columns: Context, Problem, Solution, and Example.

5.2 Recommendation 1: Improving Timeliness

The **timeliness** data quality dimension refers to the update between a change in a real-world state and the resulting modification of the state of the information system, as presented in subsection 3.3. Thus, to improve this dimension, there should be a field to inform the duration of the disease.

There are many diseases and limitations that are acquired or even diagnosed throughout life. Others are diseases that present themselves at birth. In addition, there are diseases that can be considered permanent or temporary. Permanent diseases are considered diseases that have no cure and will be present throughout the student's life. For example, multiple sclerosis is an autoimmune disease that has no cure yet. Cerebral palsy, physical paralysis and blindness are also examples of permanent diseases. According to Frausin (2022), permanent disability concerns motor and cognitive impairment.

Temporary diseases are illnesses that the student may have developed throughout his or her life, but that only manifest for a short period of time. Also according to Frausin (2022), temporary disability is related to a short-term injury. As an example, a person in a car accident may lose leg movement due to the trauma caused by the accident, but this loss may not be permanent. After a certain period of time, the student becomes physically incapable of moving around. Knowing this temporality of the disease can be extremely helpful so that adaptations and initiatives can be developed to assist in the educational development of these students.

An update should be made as soon as an diseases or disability is diagnosed for a particular student. In addition to registering the disease and information related to it, it is also necessary to identify the start date of the diagnosis. If so, the end date of the disease or limitation must also be filled in. Even if the illness is temporary, this data should not be deleted when the illness or limitation has been cured, but updated. Thus, the entire medical history of diseases can be consulted.

Along with this information, it is recommended to have a Boolean-type field that is used as a flag, to indicate whether a certain disease or limitation is active or not in that period. With this information, it becomes possible to provide accessibility in a more assertive way in educational institutions.

To improve the timeliness data quality dimension, it is also important to include a field informing whether the diagnosis is official, i.e., to identify whether a given illness or limitation is a formal diagnosis or an empirical observation, such as one made by a principal, secretary, or teacher. Moreover, it is critical to record which person is responsible for issuing the information whether the diagnosis is official or not. And it includes the occupation of that person.

It is not always possible to diagnose certain diseases or disorders in students immediately or early. In some cases, these limitations begin to manifest later in a student’s life. However, some symptoms and behaviors can be observed early on, and may contribute to later diagnoses. This is because many behaviors that are observed in childhood have a very strong influence on diagnoses that are observed at a much later age (GRILLO; SILVA, 2004).

Furthermore, it would be necessary to keep a behavioral history of the students, which was observed by education professionals, and that could help with future diagnoses. This information, in addition to contributing to a quicker diagnosis, would facilitate the traceability of this information. This is because the education professional could follow considerable data about the behavioral changes and evolution of a given student.

These data must be temporal, i.e., it must be updated according to the student's development. Because in addition to helping with possible diagnoses, it allows education professionals to stay up to date. This enables them to adapt classes and school activities to include all students, especially those who have already identified limitations.

Table 3 shows a summary of the Timeliness dimension recommendations.

Table 3 - Summary of the Timeliness dimension

Timeliness	
Context	There are diseases that are acquired throughout life and others that are already born with the person.
Problem	Disabilities or conditions can be temporary or permanent. In the former case, once the bearer has recovered or cured, this information may become outdated, generating negative effects on other quality dimensions.
Solution	Include the fields: temporary_disease, start_date, end_date, flag_active, oficial_diagnosis, diagnostician, diagnostician_profession, history_diagnosis
Example	temporary_disease=True, start_date="2023-01-01", end_date="2023-01-20", flag_active=True, oficial_diagnosis=False, diagnostician="Ana Luiza", diagnostician_profession="Teacher", history_diagnosis="Student began to experience difficulty in writing in January 2020."

Source: The author, 2023.

5.3 Recommendation 2: Improving Accuracy

The data quality dimension of accuracy verifies that the modeled data is in line with the real world, i.e., it verifies that the information is accurate enough and close enough to reality 3.3. Thus, to improve the data quality dimension of accuracy, it is recommended to include a field that informs the Classification of Diseases. By using this field, it possible to determine with accuracy what disease or limitation the student has, according to an official classification.

Listing all the existing diseases would be extremely extensive and beyond the scope of this work, therefore the International Classification of Diseases (ICD) by the World Health Organization, as presented in subsection 1.2, should be used.

The main advantage of this classification is that it has an alphanumeric code for all existing diseases around the world. This code can be understood by any doctor, i.e., it is considered a universal language in medicine. As this information about the disease/limitation that the student has is filled in according to an international and reliable source, there is no need for this data to be filled in manually, in which the student needs to write in an open field what the disease he has.

Thus, it is recommended to include two fields in the base table belonging to the unified database of PWD students. One of the fields would be the alphanumeric code of the disease categorized according to the international classification of diseases. Since this classification undergoes updates over the years, it is necessary to include another field to identify the source of this information, i.e., in which version of the ICD that data is found. So, to organize the database, the disease code field is a foreign key inside the main table. This foreign key will relate to other tables that will have a list of diseases and codes, one table for each version. Nevertheless, with these two fields, it possible to cross-reference normalized data. In addition, we able to obtain the exact name of the disease that a certain student has.

Table 4 shows a summary of the Accuracy dimension recommendations.

Table 4 - Summary of the Accuracy dimension

Accuracy	
Context	As different diseases can have the same names, it is important to have a universal nomenclature.
Problem	There are many identical names for the different diseases.
Solution	Include the fields: ICD_classification_fk, ICD_version
Example	ICD_classification_fk=10, ICD_version="ICD-11"

Source: The author, 2023.

5.4 Recommendation 3: Improving Duplication

The data quality dimension of duplication relates to the redundancy of records or attributes, according to subsection 3.3. Thus, to improve the data quality dimension of duplication, it is recommended to include some basic personal data to help guarantee uniqueness in the database. It is through these basic data that we able to ensure the uniqueness of each student's information.

For this basic personal data, we suggest the inclusion of the following attributes:

- CPF (Individual registration) or SSN (Social Security Number)
- Name
- Cell phone
- Email
- Date of Birth
- Gender
- Address - Country
- Address - State
- Address - City
- Address – Neighborhood

These attributes, despite being common to all students, are extremely important to help ensure the uniqueness of the database of students with disabilities. Furthermore, this information can be used for various analyses and follow-ups. From them, it is possible to segment the base and help in the elaboration of public policies for students with disabilities in each region of the country.

It is pertinent to note that the attributes listed can be considered sensitive data under the LGBD law, as presented in subsection 2.1.2. So it is very critical to consider all the requirements made by law and respect the confidentiality of this data.

In order to avoid redundant data in the database, it is also recommended to use the CPF as the primary key. This issue is important because if we have more than one piece of information for the same student, it is difficult to know which one is correct, which can even lead to some damage in the teaching of this student.

Another way to contribute to improving the data quality dimension of duplication is to use unique identifiers (UI‘s). For all fields that have fixed and pre-established information, such as International Classification of Diseases, presented in the subsection 5.3, it is recommended to use these UI‘s. Using them, every time you reference an attribute, just refer to its UI, and no more duplication was necessary. Moreover, it eliminates the risk of the same attribute value being written in different ways, since it written only once and always referenced by its UI.

Table 5 shows a summary of the Duplication dimension recommendations.

Table 5 - Summary of the Duplication dimension

Duplication	
Context	In a database with lots of different and incompatible information about each student, it is hard to know which information should be used to improve that student’s education.
Problem	Redundant data can generate wrong information.
Solution	Include the fields: cpf, name, cell_phone, email, date_birth, gender, address_city, address_country, address_state, address_neighborhood
Example	cpf=136.333.222-09, name=“Maria”, cell_phone=“21 99999-8888”, email=“maria.silva@gmail.com”, date_birth=“24/02/2001”, gender=“F”, address_city=“Rio de Janeiro”, address_country=“Brazil”, address_state=“Rio de Janeiro”, address_neighborhood=“Laranjeiras”

Source: The author, 2023.

5.5 Recommendation 4: Improving Coverage

According to subsection 3.3, coverage refers to the part of real-world entities that are represented in the data. Thus, if only students who have an official diagnosis were included in the base, this coverage of students would be lower. So, also considering students who have not yet had an official diagnosis, but who demonstrate behaviors that may negatively influence their student performance, coverage would be more extensive and more accurate. Increasing the quality of the database according to this dimension.

The attributes presented in the subsection 5.2, which are responsible for indicating whether a diagnosis is official or not, are critical for improving the data quality dimension of coverage. Furthermore, the behavioral history of the students, which is provided in the subsection 3.3, is also needed to improve the quality of this dimension. With a complete history of student behavior, it possible to increase the coverage of information that the database has. It brings richer content for education specialists to work on.

Another key factor contributing to the improvement of the data quality dimension of coverage is the use of the international classification of diseases. With the use of this internationally standardized database, no disease or limitation already discovered will left out of the database. Thus, the database will provide wide coverage in relation to the variety of diseases and existing limitations in all countries. In addition, this classification also increases the coverage of professionals who able to use this database. Because diseases are classified with standardized codes, any health professional can understand which disease or limitation that data is referring to.

Table 6 shows a summary of the Coverage dimension recommendations.

Table 6 - Summary of the Coverage dimension

Coverage	
Context	The symptoms and difficulties a student may experience can manifest themselves early in their lives, even before a diagnosis has beenmade.

Problem	Official diagnoses are not available to all students.
Solution	Include the fields: oficial_diagnosis, history_diagnosis, ICD_classification_fk, ICD_version
Example	oficial_diagnosis=True, history_diagnosis="Student began to experience difficulty in writing in January 2020.", ICD_classification_fk=10, ICD_version="ICD-11"

Source: The author, 2023.

5.6 Recommendation 5: Improving Completeness

According to subsection 3.3, completeness refers to the degree to which values are present in a data collection. In other words, it refers to the presence of empty values. As a result, it is recommended that some fields be mandatory, i.e., that some fields cannot be NULL.

If all fields in the database are mandatory, we cannot guarantee 100% completion, as not all fields need to be filled in. As the database of students with disabilities that is being discussed in this work will contemplate both students with a diagnosis that has already been officially determined and students who still do not have a closed diagnosis, but who have behavior that may somehow impair their school performance, it is necessary to provide flexibility in certain fields. In the next paragraphs, the fields that must be filled in will presented to improve the completeness of the data.

The fields that have basic personal information, as presented in subsection 5.4, must be mandatory. Additionally, without these fields we would not be able to guarantee the uniqueness of each student in the database, regardless of whether they are disabled or not.

The tables that have the international codes of the diseases, as seen in subsection 1.2, are dimensions that must be mandatory. Thus, every disease code must be linked to some disease, i.e., these dimensions must not have any empty fields. However, in the main database referencing these dimensions, the foreign key field may be empty. As the database may have students who do not yet have an official diagnosis, not always a student will have a disease code linked to him and thus this foreign key field will be empty.

The Boolean-type field that indicates whether a certain disease or limitation is active or not, as presented in subsection 5.2, must be filled in mandatory. Even if a certain student's illness has not yet been officially diagnosed, it is important to know whether such limitation or behavior may be affecting that student's school performance. Thus, this information becomes indispensable in this possible database of students with disabilities. The unique identifier fields proposed in subsection 5.4 must also be mandatory. All these keys created must be related to some data, i.e., there cannot be a key that does not have any unique data linked to it and, as a consequence, no key must be created without a purpose.

The behavioral history, presented in subsection 5.2, must be mandatory. Regardless of whether the student has already been officially diagnosed with one or more than one diseases, he must have his behavioral history always completed and updated. With this history, future diagnoses can be discovered, in addition to providing inputs so that ed- ucation professionals can adapt classes and school day-to-day to be able to include all students, especially those who need some more specific treatment.

Table 7 shows a summary of the Completeness dimension recommendations.

Table 7 - Summary of the Completeness dimension

Completeness	
Context	Some information that should be mandatory about PWD students, because without it, some treatments and actions cannot be developed.
Problem	Some information is critical to help with the educational development of PWD students.
Solution	Fields that must be mandatory: cpf, name, cell_phone, email, date_birth, gender, address_city, address_country, address_state, address_neighborhood, flag_active, oficial_diagnosis, history_diagnosis

Example	cpf=136.333.222-09, name="Maria", cell_phone="21 99999-8888", email="maria.silva@gmail.com", date_birth="24/02/2001", gender="F", address_city="Rio de Janeiro", address_country="Brazil", address_state="Rio de Janeiro", address_neighborhood="Laranjeiras", flag_active=True, oficial_diagnosis=True, history_diagnosis="Student began to experience difficulty in writing in January 2020."

Source: The author, 2023.

5.7 Recommendation 6: Improving Conformity

The conformity dimension is responsible for determining whether the data conforms to a specific format, such as data type or size, for example, as presented in subsection 3.3. Thus, to improve the quality dimension of compliance data, it is recommended that all fields that are possible to predetermine their values, be predetermined. For example, in personal data fields such as name and CPF, this rule does not apply because they are infinite, unique and private values for each student. But fields like gender and address (Country, State, City and Neighborhood) may already have all their possible values listed in different dimensions.

As students register in the main database, to fill in those fields that already have their predetermined values, it is enough to reference them with their respective foreign keys of the dimensions, without having to type them again. Thus, for these fields, the data format will always conform to the specification.

For the disease classification attribute, presented in the subsection 1.2 as international disease codes, the values of these fields are already predetermined, also contributing to the improvement of data compliance.

It is important to consider the type of disabilities as a new attribute with predetermined values that may contribute to improving conformity. There are thousands of types of disability in the world. Listing all of them is beyond the scope of this work. Upon evaluating the literature, as presented in subsection 1.3, we will classify the types of diseases to recommend the inclusion of this attribute as an imperative factor to consider. It is pertinent to point out that each student may have more than one type of disease, as it is possible for him to have several diseases simultaneously. Based on this information it will possible to derive many relevant analyses.

In the area of Education, this classification of disease types is crucial because different initiatives can be developed based on it. For example, for students with physical disease, in specific physical education classes, it is necessary to make adaptations for these students. A traditional volleyball class cannot be taught to all students equally. Thus, education professionals equipped with information on the types of diseases their students have will able to adapt classes to be able to include all students.

For students with cognitive issues, it will often be necessary to adapt the way in which assessments are carried out. For certain diseases, very long traditional tests can be harmful to the student's development. Having this information about what types of disease each student has allows adaptations that would not have been possible before. In addition, if the teaching unit requires students to fill out forms that are not adapted for these students, it is possible that the quality of the information entered will affected. So, certain mechanisms should be created to help with these issues, such as creating fields in forms or tests with predetermined values.

Table 8 shows a summary of the Conformity dimension recommendations.

Table 8 - Summary of the Conformity dimension

Conformity	
Context	Standardizing certain information that can be previously defined helps in ensuring data conformity.
Problem	The same information can be registered in different ways.
Solution	Fields that should have their predetermined values: ICD_classification_fk, ICD_version, types_disabilities, gender, address_city, address_country, address_state, address_neighborhood
Example	ICD_classification_fk=10, ICD_version="ICD-11", types_disabilities="Cognitive Disorder", gender="F", address_city="Rio de Janeiro", address_country="Brazil", address_state="Rio de Janeiro", address_neighborhood="Laranjeiras"

Source: The author, 2023.

5.8 Recommendation 7: Improving Consistency

The consistency dimension is responsible for determining whether the data in all systems reflect the same information and are synchronized with each other throughout the database, as presented in subsection 3.3. Thus, to improve the consistency data quality dimension, it is recommended to use identifying fields, such as primary and foreign keys, for example.

Subsection 1.2, shows the international disease codes and the versions of the ICD codes they are linked,they are great fields for improving consistency. Because with them it possible to reference disease classifications uniquely, ensuring that this information is consistent throughout the database.

Other recommended fields to help improve this dimension are the UI‘s fields explained in the subsection 5.4. In addition to helping to improve the dimensions of duplicity and completeness, they also contribute to maintaining data consistency across the entire database. Instead of always having to write that information several times, which may cause an error or different writing, the pre-established UI‘s referenced, helping to improve consistency.

Other attributes also recommended to help improve consistency are personal data attributes, such as name, CPF, and email. With these fields, it is possible to identify each student uniquely and use them to reference each student, thus ensuring that the data of each student consistent throughout the database, i.e., that all data referring to those students reflect the same information in the whole base.

More specifically for the PWD student scenario, helping to ensure data consistency extends beyond verifying that certain information is being reflected in the same way across different tables and databases. In this context, it is imperative to use some mechanism that allows periodic confirmation of certain information. For example, if a particular student reports that he has a permanent illness that affects his academic development, this information needs to be revisited. This is because even permanent illnesses can bring new complications and symptoms that affect that student’s academic performance. So to ensure data consistency, periodic verification becomes essential.

PWD students must have their data periodically verified according to their specific diseases or limitations. A student with a hearing impairment, for example, must confirm information through text or visual means. In the case of visually impaired students, data can be verified through interviews. Each limitation must be validated in some way. This periodic verification, in addition to contributing to the consistency dimension, also helps to improve the timeliness data quality dimension.

Another way to improve the consistency of PWD students' data would be to bring some information already made available to public agencies responsible for issuing citizens with documents such as Identity Card (RG), Passport, and CPF, among others. In addition to the differentiated identity card, which is an identity card containing the indication of what type of disability that student has, the government also provides a PWD card (Governo do Estado do Rio de Janeiro, 2023). This card already has some important information for students, such as Type of Disability, the International Code of Disease (ICD), description of medicines for continuous use, types of allergies, and personal contacts for possible emergencies. In addition, the information available through the RNDS platform, presented in subsection 1.4 would help in improving consistency.

People with disabilities such as physical, mental, visual, auditory, intellectual, and/or autism can request this card. This way, the consistency of data on PWD students is improved by bringing in previously registered information.

In addition, it is necessary to carry out periodic checks against these public databases, which be used to bring information. Because, as these databases are updated, they don't reflect the data in the unified database. Thus, with these periodic verification, a database must be flexible to changes.

Table 9 shows a summary of the Consistency dimension recommendations.

Table 9 - Summary of the Consistency dimension

Consistency	
Context	The data from the different databases must be synchronized with each other to perform consistency analyses on PWD students.
Problem	Data may not be synchronized across the entire database and maynot reflect the same information.

Solution	Include the fields: ICD_classification_fk, ICD_version, cpf, name, email. Create an information verification mechanism and bring data already registered in the government database.
Example	ICD_classification_fk=10, ICD_version="ICD-11", cpf=136.333.222-09, name="Maria", email="maria.silva@gmail.com"

Source: The author, 2023.

5.9 Recommendation 8: Improving Integrity

The integrity dimension concerns the ability to link related records to each other, i.e., it is responsible for verifying accuracy based on data relationships, as presented in subsection 3.3. Integrity can also be understood as the sum of three other dimensions of data quality: accuracy, completeness and consistency (REIS, 2020). As long as the data is of quality in these 3 dimensions, it can be stored, consulted and used with confidence, thus being integrity data (REIS, 2020).

To improve the quality of integrity data, it is recommended to apply certain storage restrictions, called integrity restrictions, related to aspects such as types of data stored, relationships between primary and foreign key columns, the possibility of having null or non-null values, and rules specific business.

The first recommended integrity is domain integrity. This restriction is related to the definition of values that are allowed in each column of the database, i.e., they are the values of the domain (REIS, 2020). In addition to the data format, check whether that field is mandatory and whether it belongs to a set of predetermined values. For the specific domain of PWD students, to guarantee this integrity, it is recommended that the fields that may already have their values already predetermined be filled in, as presented in the subsection 5.7, such as:

- Gender
- Address (Country, State, City and Neighborhood)
- International disease code
- Types of disabilities

Furthermore, the subsection 5.6 addresses the issue of which fields should be mandatory or not, contributing to the issue of null values in the database.

The second recommended integrity is referential integrity. This restriction is responsible for securing that the values of a column in a table are valid based on the values in another related table (REIS, 2020). To guarantee this restriction, it is recommended to use primary and foreign keys and unique identifiers (UI‘s) as explained in subsections 5.8 and 5.4. In addition, considering the specific situation of PWD students, this restriction must also ensure that all listed diseases and limitations contemplated in the database must refer to the international disease codes presented in the subsection 5.3.

The third recommended integrity is null integrity. This restriction is responsible for determining whether filling in a column is mandatory or optional (REIS, 2020). A primary key column, for example, must always have data inserted, and can never be empty, for any record. To guarantee this restriction, it is necessary to follow the recommendations made in the subsection 5.6, which lists the fields that must be mandatory for the context of PWD student data. Thus, based on the explanations in the subsection 5.6, the following fields cannot have null values:

- Fields that contain basic personal information, as presented in the subsection 5.4.
- The tables that have the international codes of the diseases, as seen in the subsec- tion 5.3.
- Boolean-type field that indicates whether a given disease or limitation is active or not, as presented in subsection 5.2.
- The unique identifier fields, proposed in the subsection 5.4.
- The behavioral history, presented in the subsection 5.2.

The last recommended integrity is key integrity. This restriction is responsible for ensuring that the values inserted in the primary key column must always be unique, not allowing repetitions of these values (REIS, 2020). Thus, it is recommended to validate that all key fields and all unique identifier fields are unique in that database.

Table 10 shows a summary of the Integrity dimension recommendations.

Table 10 - Summary of the Integrity dimension

Integrity

Context	As the information will be stored in different tables, there must be some mechanism that guarantees the integrity between them.
Problem	Information about PWD students needs to be correctly related to each other.
Solution	Follow the rules of integrity: domain integrity, referencial integrity, null integrity
Example	Domain integrity –> The fild Types of disabilities can only contain the following values: Cognitive Disorders, Psychological and Physical

Source: The author, 2023.

As it was possible to see throughout the chapter, there are some recommendations that can be used as a guide in building a unique database for PWD students. The goal of these recommendations is not just to build any database, but to build a quality database.

6 PROPOSAL EVALUATION

The proposal of this dissertation - a set of recommendations aimed at guaranteeing the quality of a database of PWD – has clearly 2 aspects. The first one is related to the technical issues of applying recommendations to build a database. The second one is the usage of a database with such characteristics.

While the first should be understood and assessed by database or system development professionals, the second is closely related to the stakeholder educational field. Therefore, the methodological approach chosen to evaluate the proposal was interviews and experts' panel.

This chapter presents how the methodology for the evaluation was chosen, the ques- tions asked in the interview and in the experts' panel, details the execution of the evaluation itself, the analysis and discussion of the results, in addition to exposing the final considerations.

6.1 Methodological approach

The methodological approach chosen in this research aimed to evaluate whether the proposal is technically feasible and if it can contribute in any way to professionals in education. The goal was is to evaluate the proposal in two different perspectives. Thus, we decided to conduct interviews with professionals from the technical and education fields to collect perceptions from a variety of perspectives.

According to Marconi & Lakatos (2003), the purpose of an interview is to get information from the interviewee about a certain topic, to investigate theories, to determine the opinion on certain facts and even to obtain opinions. In a semi-structured interview, the researcher formulates questions to the interviewee, following a previously established script, in order to acquire data that are of interest to the research (FREITAS et al., 2022; MARCONI; LAKATOS, 2003). During the the conduction of the interview, the researcher follows the previously established questions, being able to add new questions if necessary.

The experts' panel technique, developed in the 1960s, allows the qualitative validation of evidence. This is, by using a group of experts on the subject, it allows the selection, judgment, and creativity of suggestions for the evaluation of a proposal (CAMPOS et al., 2010). This technique is a structured variant of the methods for small group discussion. In this sense, by preventing one person from dominating the debate, it encourages passive participation (CAMPOS et al., 2010).

Once the interview technique was chosen, we designed a semi-structured interview because the idea was to follow a script of questions, but without imposing limitations on the conversation, so that we could ask other questions that emerge during the interviews. Another question that arose was whether the interviews would be carried out physically or remotely. Analyzing the current scenario, where with the pandemic remote meetings and work have become more common and efficient, we opted for the remote interviews.

Regarding the conduction of the interviews, the protocol followed was: the interview started with a brief slide presentation about the dissertation proposal and then the script with questions was asked. For this presentation two ideas emerged, the first was to record a video presenting the proposal and the basic content for the questions, where the interviewees should watch it beforehand. The second idea would be that in each interview the presentation would be made in person and in real time for each interviewee. We chose to follow the second idea, because with it we could guarantee that the interviewees would be watching the presentation, and if any questions arose, they could be solved on the spot, without running the risk of forgetting.

Delving further into how to conduct the interviews, it was decided to switch from interviews to experts' panel during the technical assessment. This experts' panel technique was chosen because of its possibility to involve different people at the same time, and also because of its ability to generate many opinions in a short period of time. In addition, it enables the technical professionals to come up with opinions that they would not have had if they had been interviewed individually. For the evaluation by education professionals it was decided to keep the interview technique, because the objective was to bring more personal feedback, involving the professional's day to day experience.

6.2 **Technical assessment**

To perform the experts' panel, 5 professionals working in the field of Database were invited: 3 data engineers and 2 data specialists. To increase the diversity of the participants, people of different genders, ages, educational levels, length of experience, and country of work were chosen. The following list summarizes this characterization data:

- Gender: 60% male and 40% female
- Age: 80% are between 25 and 34 years old, and 20% are between 45 and 54 years old
- Education: 80% computer science, and 20% information systems
- Education level: 40% post-graduate, 40% master's, and 20% undergraduate
- Time of experience: 60% between 5 and 12 years, and 40% between 14 and 20 years
- Practice area: 60% data engineering, and 40% data analysis
- Country of work: 60% Brazil and 40% Germany

In addition, some questions were prepared previously, with the objective of guiding the experts' panel, but without restricting it. The questions and topics that were addressed during the experts' panel will be presented below:

1. Is the application of these patterns to build this database feasible at the implementation level? Can you identify some techniques that could be used?
2. Does the use of the mentioned dimensions contribute in any way to the construction of the database?
3. Are there other technical issues that should have been taken into account in the proposal? What are they?
4. Did you miss any restrictions? What is it?
5. Did you miss any specifications regarding the base? Which specifications? What adjustments would you make?
6. Do you have any restrictions that you would not include in the proposal? Or any dimension that you don't think is necessary/important?
7. Future works

8. What positive points can you see in the proposal, thinking on the technical side?

Before starting the experts' panel, a slide presentation was made with a summary of the proposal of this work, as presented in Appendix A. The purpose of this brief presentation was to introduce the topic studied and explain the deliverables and objectives of the dissertation. After that, the debate started, leaving an open time for the participants to give their opinions about the presentation and the proposed theme.

The debate occurred in a very positive way, where all participants felt comfortable exposing their opinions and ideas. However, at some moments it was necessary to interrupt and remind the participants what the objective of the proposal was, because as it is a broad theme, at some moments many ideas and opinions emerged that diverged from the objective.

The next subsections present the discussions for each question. In addition, to facilitate the understanding of the discussions and to guarantee the anonymity of the experts, each participant will be referenced by a code: P1, P2, P3, P4, and P5.

6.2.1 Question one

Regarding the first question, which concerns the feasibility of implementation, the five participants agreed that implementation of this proposal is feasible. And based on this, some suggestions for implementation techniques have been identified. Participants P3 and P2 brought the most opinions to the discussion, as will be presented in the following.

According to participant P3, since the proposed database recommendation is to contain only PWD students, should a student cease to be considered PWD or be cured of a certain disease or limitation, he suggested that student's data be logically deleted. In this way, the data would not be lost and could be consulted for future analysis. This proposed logical deletion would be the creation of a flag to control which students are PWD and which are not, as he mentions below:

“This is related to what I said about removing non-PWD students from the base. Not physically delete them, but delete them logically. Use a flag to do this logical deletion.”

The other four participants agreed with this implementation technique suggested by participant P3.

In the proposal there is already a suggestion to include a flag field to indicate whether a certain disease is currently manifesting itself or not. The flag, however, is associated with each student’s disease. The flag for logical deletion would have the objective of indicating that the student no longer has any disease. So it would be viewed by students, and not by disease.

In addition, this suggestion of the logical deletion flag contributes to data quality dimensions timeliness and coverage. In the dimension of timeliness, it contributes to updating the data, as it can identify which students are currently PCD and which are no longer considered PCD. In the dimension of coverage, it increases students’ comprehensiveness and coverage.

Another technique proposed by P3 was the creation of N to N relationships in the database. This technique was suggested after a questioning by P1:

“You talked about illness registration. Considering that one person has two disabilities, would you duplicate this in the base, how would you handle this kind of situation? ”

P3 argues that, this is because one student may have many diseases, and one disease may be associated with more than one student, as he mentions below:

“I see it as an N to N relationship. Because one student can be related to many diseases, and many diseases to many students.”

In addition, the technique of versioning the data was also suggested by P2. Let’s say a PWD student does not have any other diseases besides their logical deletion. In that case, it would be interesting to know the date when the change occurred, as well as the summer when that event took place. For if the same student has some illness or limitation again, that history would not be lost. The following shows this P2 argument:

“When you say that the base has to be flexible, i.e., it can have a disease and get better and then have it again, I wanted to suggest something about versioning. It would be to version the activities and lifespan in the base.”

The other four participants also agreed with this implementation technique suggested by participant P3.

This suggestion of the versioning technique contributes to the data quality dimension of timeliness. This is because it makes the database dynamic and helps ensure it is always up to date. It also contributes to the completeness dimension, because by requiring this versioning information to be mandatory, we will have a more complete database.

Another implementation technique identified, was thinking about updating the static dimensions, such as a dimension that contains all diseases and ICD codes. Participant P2 brought the following suggestion:

“Maybe to update the dimensions it would be nice to have some kind of update mechanism. This could be an automated crawler that surfs the internet once a month, and updates the dimensions automatically. And then the teacher, for example, when he or she fills out the form, would have a field to choose which disease. It can be any technology, but use the crawler concept to update this data.”

P2 argues that it would be interesting to use the crawler concept to perform this update. The idea would be to have some mechanism that automatically, from time to time, consults the Internet in search of new data to populate that dimension.

To complement this suggestion, P5 brings the idea of using possible APIs:

“If this international disease registry already exists you must have some API ready to consume this data.”

It would contribute to data quality dimensions of consistency, timeliness, and completeness of data. In terms of consistency and timeliness, it would ensure that data from the static dimensions is always synchronized with the main sources of information and always up-to-date. On the other hand, in the completeness dimension, by making this information mandatory, the database becomes more complete.

Another technical issue was mentioned by P3:

“I wanted to make one more technical suggestion. What about the backups of this data? It is very important to have backups of this data. If something happens, to have a way to recover it.”

Participants P4 and P5 agreed that this technique of using data backups is extremely important. And finally, another technique suggested by P2 was to use the user experience of the database to help build the database:

"Maybe you can use the experience of the user who uses this data. For example, when I tried to buy things at Casas Bahia, the clerk left several fields blank. This is because he was in a hurry and didn't want to spend a lot of time filling out the form. So thinking about the teacher, this set of checks ensured that the teacher could fill in the data with more quality. Making the filling out experience pleasant. Because the teacher is already overburdened with very limited time. So when modeling, it is nice to consider user experience. Define fixed formats for fields with predefined information, for example. Thinking about coverage, there could already be a list of symptoms. The professional would ask the person responsible if the child has those symptoms before marking it as a checklist. User experience would be based on data."

User experience, as suggested, would contribute to data quality dimension conformity. With the standardization of certain information that can be predefined based on user experience, it would help ensure the conformity dimension.

6.2.2 Question two

Regarding the second question, the participants were asked if the use of the mentioned dimensions contributed to the construction of the database. Dimensions are relevant to recommendations, and the five participants agree they should be used. To contribute to this statement the participant P4 gave the following statement:

"It is very valid to use the dimensions, because they are concepts of data quality, there is no way we can use them."

In addition, the P3 argued that most of these dimensions will be applied at design coding time, i.e. they will be heavily used at database creation time. The participant P1 brought up the idea that not all dimensions will be applied to all tables in the base:

"I think it would help. Maybe, for example, the case of duplicity does not apply to all tables, but all tables that you are going to build, like I was thinking here of a modeling and maybe, for example, duplicity does not apply to a disability registration table, so you would have to duplicate the data there to cross-reference it with your student table. Students could have more than one disease."

That is, if a student can have more than one disease and limitation, it will be necessary to make an N to N relationship. This will mean that the student will have more than one row in the table. The duplicity would not be in the data, but in the concept itself.

6.2.3 Question three

In the third question, it was asked if there were other technical issues that should have been taken into account in the proposal. The participant P4 felt that it was necessary to have minimal modeling to facilitate the understanding of the work proposal. Having a visual representation would facilitate understanding:

"It gets a little confusing to discuss these issues without modeling, because it becomes too abstract."

The other four participants agreed with this question. According to him, 24% of the Brazilian population has some kind of disability. So there are about 40 million people in the database, which is not a very large amount for a database. But thinking about technology, it would be interesting to calculate the number of PWD students and multiply it by the number of possible diseases that affect them. Thus, we have an approximate measure of the minimum amount of records, serving as input for the choice of technologies to be used.

This suggestion to make recommendations on technical limitations contributes to ensuring data quality dimensions of coverage and completeness. With inputs to the choice of technologies to be used, the database can be built in a flexible and scalable way. In this way, the data coverage can be expanded and more flexible, as well as complete.

Another interesting point, but beyond the scope of this paper, is reported by P2 in the following speech:

"We missed talking about data encryption, some security issues. Thinking about the day to day of Brazilian education, many times teachers don't have resources, don't have the internet."

As this question would be more related to the development itself of the database, we did not delve into this subject during experts' panel.

6.2.4 Question four

When the discussion about the fourth question started, the participants were asked if they missed any restrictions in the recommendations. Two significant points were raised. The first was about the basis being limited to PWD students. The participant P1 felt that the database could be for all students and have some mechanism to identify which students are PWD and which are not. Participant P3 agreed with this point.

Then, it was explained that in order to restrict the scope of the master’s thesis, only PWD students were chosen. Expansion to other students could be done in future work.

The second point was regarding the word “flexible”, brought by participant P5. He commented that this word sounds like it can be deleted and overwritten:

“When you talk about flexible tables, sometimes flexibility makes me afraid. Because if it is flexible it means such an update is in place and we would lose the history.”

So it was suggested to be careful when using it. In contrast, the P3 also suggested to make it clearer that the data would be deleted only logically, keeping a history of all the information.

These two proposed recommendations help ensure the quality of coverage and timeliness of data. As a result, including all students, even those who are not PWD, would increase the database coverage. In addition, the recommendations for a flexible database contribute to the timeliness dimension, as it allows the database to stay up-to-date and flexible to change.

6.2.5 Question five

For the discussion of the fifth question, the participants were asked if they missed any specifications regarding the base. Based on this, what adjustments they would make. This question was well discussed by the participants and many opinions were brought up for discussion.

The participant P3 said that he missed a link between the student and his legal guardian:

"Sorry to interrupt, but I missed that connection between the student and his legal guardian. Have some kinship information as well."

So it would be interesting to include additional information such as legal guardian, how related, and even an emergency contact. In order to ensure completeness of data quality, this information should be mandatory in the base.

In addition, the P4 suggested that besides behavioral history, the family history field would be interesting. This new field would be filled in with the student's behavior at home, observed by the family, especially by the parents. Because many times the student's behavior at home can bring a lot of interesting information for professionals to take into consideration when making a diagnosis, as P4 mentions below:

"Because there could be an initial input that would be, I am responsible, I have my son, I know he has the disease right."

This new field would help in ensuring the quality dimension of timeliness data, as it is information that changes over time, and every day various symptoms and behaviors may arise that must be constantly updated and entered into the database.

The third topic was related to those responsible for filling out the database. The participants P1 and P2 felt that there was a lack of any recommendation that specified which people should be responsible for filling out the database. The suggestion is that this responsibility should be divided among parents, teachers, and doctors. And the initial filling in could be done by the secretary of education, as soon as the student enters the school network.

Recommending who is responsible for completing the information contributes to ensuring data quality dimensions of completeness and coverage. By choosing the right people for this task, it is possible to ensure that the data will be filled in from more than one source of information. In addition, the data will be filled in more accurately and completely.

Another recommendation that the participant P2 missed was the specification of who could have access to the database. Because it is very sensitive data, it is dangerous that anyone can use it:

"Who would be the people that would have access to this database? Because I am afraid that some people will use this database for evil purposes, for example, health insurance companies finding out that the family has disabled children and charging more for this.”

So it was suggested by the other participants that only education professionals who had direct contact with a particular student could access the data related to that student. Or, if the educational institution had a committee dedicated to PWD students, they could have access to the database to make analyses and create improvement initiatives. Regarding doctors having access to this database, it was suggested that only doctors authorized by the students’ parents could have access to the students’ data.

It contributes to the LGPD guarantee mentioned in subsection 2.1.2 by recommending who could have access to the information. This data must be accessed according to the law, and all personal and sensitive data rules must be followed.

At the end of the discussion of this question, it was suggested by P1 and P4 that more analytical recommendations be made, as shown in the speech of the participant P1 below: “It would be nice to also bring in a more analytical look. What do I want to answer later on? How do I want this data to be analyzed? What answers do I intend to give?

That could help in building the database.”

6.2.6 Question six

In the discussion of the sixth question, it was asked if there were any restrictions that they would not include in the proposal. In addition, it was also asked if there were any dimensions that they would not include in the proposal. Regarding the dimensions, the five participants agreed that all of them should continue to be included in the proposal.

However, one of the fields recommended in the proposal, the "gender" field, was suggested to be removed from the base, by participant P5. The reason was explained in the following speech:

“Since it is an open field, we would have no way to guarantee the domain integrity of it. And furthermore, since it’s information that doesn’t have a direct impact with diseases and limitations, it might be interesting to take it out of the database. Besides he is a very sensitive and controversial piece of information.”

Thus, by removing the "gender" field, it is possible to contribute to the data quality dimension of data integrity. Because, as it is an open field, it is difficult to guarantee the integrity of the domain. As it is not an essential field for the database, removing this information is interesting.

6.2.7 Question seven

During the proposal discussion, the participants brought up 3 ideas that, as they are outside the scope of this dissertation, but will be addressed as possible future work. The first idea, which was placed by P1, was to include information about the schools, to know if the schools have any specialized services for PWD students. In addition, an exclusive table could be created with this information, such as dates of inspections of the schools, grades that that school received for certain criteria, among other information. The goal would be to have inputs so that the schools could be evaluated in relation to being prepared to receive PWD students, and identify what improvements need to be made in relation to this.

This new information contributes mainly to the quality dimension of coverage data, as it increases the scope of analysis in relation to the database. The amount of contexts and information to be analyzed increases.

The second suggestion was placed by P5: Future work would be, based on the behavioral history field, to apply artificial intelligence such as NLP to derive information from this field. For example, from certain symptom patterns, we recommend possible diagnoses for that student. Finally, it was recommended that all students be included in the database, not just PWD students.

6.2.8 Question eight

To close the experts' panel, the participants were asked which positive points they could list. The first positive, which was placed by P3, was that the recommendations of this database will make it possible to build a database that will serve, in the future, as input for various analyses. Based on this data, PWD students will have a very high visibility, and it will provide material for analyses that may not yet exist on this subject. All the other four participants agreed with this point.

Another positive point, which was placed by P5, was in relation to the size and expansion that this work may have in the future:

"This can be the starting point of many things. We started discussing the base of PWD students and then we started talking about the school evaluation. And how do we identify the responsible parties? There is a great chance that this registry will grow a lot, so you are creating a small pile of sand that could be a giant desert in a little while."

Finally, the participant P1 brought up a positive point regarding future planning: "This proposal will help a lot in planning as well. Sometimes there are places available in certain places that don't have so many students with intellectual disabilities, and in other places there are more. So it would be possible to allocate these resources to places that need them the most."

6.3 **Educational assessment**

In order to evaluate the proposal in the area of education, we interviewed a pedagogue who works in public education. The choice criterion used was that besides having more than 30 years of experience in the area, she works in a public school, and more than 80% of the enrollment of students in the country are from the public network (MATTOS; FARIA, 2022). The following is a summary of her characterization data:

- Gender: Female
- Age: Between 45 and 54 years old
- Education: Pedagogy
- Education level: PhD

- Time of experience: 33 years
- Practice area: Public Education
- Country of work: Brazil

Some questions were prepared previously, with the objective of guiding the interview, but without restricting it. The questions that were addressed during the interview are presented below:

1. Following this suggested pattern would help education professionals in what way? What actions could you do differently with this information in hand?
2. What do you think about the idea of having a behavioral history of students? In what ways could this be used to adapt teaching for that student?
3. How would standardizing the names of diseases contribute to your analysis of PWD students? How would your work be impacted by this standard?
4. What is the importance of considering students who do not yet have a defined official diagnosis? Are there preventative measures that can be taken to help PWD students?
5. How significant is it to have this information centralized in one place? How would this contribute to the daily life of an education professional?
6. Did you miss any relevant information? What is it?
7. What do you think should be done to improve this proposal? What adjustments would you make?
8. Is there unnecessary information? Is there any information that is not contributing to improving the study condition of PWD students?

Before starting the interview, a slide presentation was made with a summary of the work proposal, as presented in appendix B. This brief presentation introduced the studied theme and explained the dissertation deliveries and objectives. After this, the interview was started, leaving open time for the interviewee to give her opinions about the presentation and the proposed theme.

In the first question, it was asked how the suggested standard would help education professionals. Also, what actions could be taken with this recommended information? Based on these questions, the interviewee said that following this suggested pattern, public policies can be developed based on this information. She mentioned that today there are 20 thousand lauded children in Rio de Janeiro. However, she estimates that more than 40 thousand children have some specificity that can compromise their school and academic development. Furthermore, this information would assist in the official diagnosis of these children, in addition to public policy.

The second question in the interview concerns the students' behavioral history. It is crucial to discuss this behavioral history with the students' parents, according to the interviewee. Because, many times, students' parents do not accept that their child may have some illness or limitation. So, the behavioral history would work as an argument to convince the students' parents that their child needs special attention. In addition, it would be very interesting if the educational professional could help feed this database with behaviors observed by him/her. After being stored, this information serves as an indicator that a particular student may need special attention, as per your comment below:

"If you have a program that helps you as an education professional, you go to the schools and feed that database. At the end, a red light turns on, drawing your attention. Not red, but a yellow light of attention. Maybe this would even be an indicator, you know. As the number of responses fills the base."

This issue related to behavioral history, as mentioned by the interviewee, contributes to ensuring data quality dimensions of timeliness, coverage and completeness. In relation to the timeliness dimension, having this field in the database allows the database to remain up to date. It also allows the database to always receive updated information as it appears in the PWD student's life. In relation to coverage, as there may or may not be students with official diagnoses, this field helps education professionals talk to parents and give inputs. This is so that their children can be diagnosed. Finally, in relation to the completeness dimension, it is critical that this field be mandatory, because without it many analyses would not be possible.

Moving forward in the interview questions, it was asked whether the standardization of disease names would contribute to education professionals' analysis regarding PWD students. This question was designed to assess whether the ICD recommendations contribute to ensuring data quality dimensions accuracy, coverage and conformity. In response, the interviewee confirmed the question, but stressed the importance of this standard:

“I think the next most helpful thing is ICD. Because I see this project as bigger. So you are already working from an international database. So you don’t have to change that. Go to ICD. I see that project in other countries. A very big thing.”

The third question was about the importance of considering students who do not yet have a defined official diagnosis. In addition, it was asked if there is any preventative work for PWD students. These questions were designed to assess whether the recommendation to include PWD students with official and unofficial diagnoses in the base would contribute to the quality assurance dimension of coverage data. The interviewee mentioned that including all these students would help parents find an official diagnosis. It would be another motivator and input to contribute to the diagnosis’s officialization.

In addition, the interviewee said she could not think of any preventive action for PWD students. Because many of the diseases and limitations that students have are already inherent. So, according to her, she does not see preventive actions even though she has all this information in the database.

Additionally, a bit further in the questions, the importance of having this information centralized in one place was questioned. The educator interviewed said that she sees a very substantial and positive impact from this centralization. She also said that data availability is a worldwide trend. Just like "sougov", which is a government platform to ensure information transparency, having this data centralized and available is a sign of citizenship:

“You ensure that this citizen is respected because he is on a unified, certified, and validated platform. This is why I am so excited about this project. Because I see this as a government project. It’s very critical, very crucial, even for granting benefits which the law provides.”

Towards the end of the interview, the interviewee was asked if she had missed any significant information. In response, she mentioned the importance of collaboration:

“Make it very clear that for this project to work, it is a collaborative project. It is a project of many hands. It is a project to help society relate fairly and honorably to all children with PWD. These children will become adults and workers.”

To conclude the interview, the interviewee was asked if she saw anything else included in the proposal. She was also asked if any recommended information was unnecessary, or even not contributing to improving PWD students' study conditions. In response, she argued that everything was extremely relevant, and that she would not take anything from the proposal.

6.4 **Discussion of results**

In this chapter, the evaluation of the proposal was described, presenting the questions used to conduct the interview and the experts' panel carried out. The execution of the evaluation and its results obtained in each analyzed sphere were also presented. One was in the technical area and the other in the educational area.

As seen at the beginning of the dissertation, the main research problem addressed in this work is the lack of inputs, through a unified database. This is so that actions and projects can be developed to increase inclusion of students with disabilities. And the biggest question to be answered is what aspects are relevant for a database to support PWD?

Analyzing the technical context, the experts were able to identify several possible implementation techniques to help in building a database with quality for students with disabilities. In addition, it was brought up during the expert's panel that the use of data quality dimensions is very critical in building such a database. It was also brought up that not all dimensions would be applied to all tables and business rules.

Looking through the educational context, having these attributes suggested through the recommendations is essential to building this database of PWD students. Also, with the possibility of creating this database, this proposal shows itself to be a much bigger project. It is a project that can reach a worldwide level to reach all PWD children from different countries. Furthermore, with this structured information many actions and projects can be developed in schools and universities to improve these students' education. For example, teachers and educators will have inputs and arguments to help those responsible for PWD students accept their children's diagnoses, which were often not accepted.

Both interviews lead to many ideas, which proves that the proposed standards make sense. For example, the suggestion to include family history information contributes to the improvement of the timeliness data quality dimension. In this way, the standard for contributing to the quality dimension of timeliness data could be increased with this idea, i.e., the proposed standard is valid but allows new information to be included to improve it.

Thus, it is possible to see that the initial question, in the beginning of this dissertation, has been answered through the recommendations presented, based on the evaluations carried out. Moreover, the results from evaluation indicates that the proposal presented is feasible.

During the evaluation some improvements to the proposal were also identified. These improvements included proposing data versioning, proposing the creation of data backups with a certain frequency, and creating some modeling to facilitate the understanding of the proposed recommendations.

Furthermore, the proposal presented in this dissertation can complement some of the related works presented in Chapter 4. For example, in Kumar et al. (2021) it is recommended that PWD related data production should be improved. They will contribute to the planning and implementation of appropriate measures to solve PWD problems. With the construction of the proposed database, it would be possible to contribute to this improved data production. This would make the present proposal complementary to this other work in the literature.

Among the limitations of the methodological approach we highlight: (i) The reduced number of participants, due to time constraints and the availability of possible participants; (ii) The difficulty of maintaining focus during evaluations. For the participants sometimes strayed from the main focus, which was to evaluate the proposal, and leaned towards the idealization of ideas based on the theme; (iii) Lack of a more visual proposal to facilitate the technical understanding of the proposal.

Despite the limitations of the evaluation presented, its execution was important (because it made it possible to get different feedbacks on the proposal, having different views on the subject) and positive (because it provided indications that the proposal achieved its objective).

The next chapter presents the conclusions of this work, the contributions of this dissertation, its limitations, and future work.

CONCLUSION

This work presented the importance of having a unified database of PWD students, resulting in several benefits for these students such as the possibility of creating government actions for PWD students, of creating inclusive lesson plans and activities, and of providing relevant information to institutions so that they can adjust based on the diversity of PWD students, among other benefits. After the literature review it was found: (i) the scarcity of approaches involving the relationship between PWD students' difficulties and the data related to them; and (ii) the absence of a formal proposal for the construction of a unified database for these students.

In this context, the described approach in this paper was proposed to assist, through recommendations, in the creation of a unified database for PWD students. To do this, the approach used the concepts of the main dimensions of data quality. This was done to ensure that this database is built with quality data.

To support the proposal elaboration, a theoretical study was done. This covered subjects such as PWD students' difficulties, the main dimensions of data quality, and Data Quality Maturity Models. In addition, related works involving data analysis of PWD students were also studied to support the proposal.

From this information, recommendations were proposed and explained throughout the proposal evaluation chapter.

An interview and experts' panel were conducted to assess whether the proposed approach was feasible and achieved its goal. The proposal was evaluated from two perspectives, one technical and one based on education. In addition, improvements were identified in the proposal. From this study there are indications that the approach is feasible and would help many students.

The main contributions of this dissertation are:

- Definition of recommendations for building a unified database of PWD students.
- Visibility of the challenges faced by PWD students, and how they can be helped in some aspects of education.

- Relationship between PWD students' limitations and data quality dimensions.
- Inputs for building a unified database of PWD students.

Limitations of this research include: (i) The recommendations made do not guarantee the totality of the data quality dimensions, i.e., the objective is not to ensure the total quality of each dimension, but rather to contribute to increasing the quality of each dimension. In subsection 5.3, for example, an improvement in data quality will be guaran- teed for the accuracy dimension, but this does not imply 100% accuracy for the database; (ii) To reduce the scope and be more specialized, the recommendations focus on PWD students, so students without disabilities or limitations were not included in the proposal; (iii) More technical issues such as recommendations of technologies to be used, analysis and studies on how to choose these technologies were also out of scope; (iv) Issues related to security, backup and data encryption were also not covered in the dissertation proposal; (v) The reduced number of participants in the proposal evaluation, due to time constraints and the availability of possible participants.

From the proposal evaluation results, some improvements and some suggestions were identified for the approach. Future work identified is listed below:

- Create a model of the proposed database.
- Implement the proposed database.
- Include recommendations regarding technology choice.
- Include information about PWD students' legal guardians.
- Expand the database to all students, removing the restriction on PWD students.

With this, it is possible to realize that many other works can be developed based on this research. These works can contribute to the improvement of PWD students' quality of life.

REFERENCES

ABREU, S. I. A. d. DISLEXIA – APRENDER A APRENDER. 116 p. Dissertação (Dissertação de Mestrado) — Escola Superior de Educação Almeida Garrett, Lisboa, Portugal, 2012.

ALCANTARA, W. et al. Desafios no uso de dados abertos conectados na educação brasileira. In: . [S.l.: s.n.], 2015.

ALVES, B. M. A influência dos alimentos no tratamento dos transtornos mentais: ansiedade, depressão e esquizofrenia. 19 p. Undergraduate (B.Sc.) Final Project — CENTRO UNIVERSITÁRIO DE BRASÍLIA (UniCEUB), Brasília, Brazil, July 2021.

(APA), A. P. A. DSM-5: Manual Diagnóstico e Estatístico de Transtornos Mentais. 2. ed. [S.l.]: Artmed Editora, 2014. 992 p.

ARALDI-FAVASSA, C.; ARMILIATO, N.; KALININE, I. Aspectos fisiológicos e psicológicos do estresse. 01 2005.

Assembléia Geral das Nações Unidas. Convenção sobre os Direitos das Pessoas com Deficiência. 2006. Disponível em: <http://www.pcdlegal.com.br/convencaoonu/wpcontent/themes/convencaoonu/downloads/ONU_Cartilha.pdf>. Acesso em: 25 de outubro de 2022.

BANDEIRA, J. et al. Dados abertos conectados para a educação. In: . [S.l.: s.n.], 2015. p. 1–24.

BATINI, C. et al. Methodologies for data quality assessment and improvement. ACM Comput. Surv., v. 41, p. 16:1–16:52, 2009.

BERMAN, J. Principles of Big Data: Preparing, Sharing, and Analyzing Complex Infor- mation. 1. ed. [S.l.]: Morgan Kaufmann Publishers, 2013. 261 p.

BIRUK, E. et al. Improving the quality of clinical coding through mapping of national classification of diseases (ncod) and international classification of disease (icd-10). In: AFRICAN JOURNALS ONLINE. African: [s.n.], 2021. (AJOL), p. 1–7.

Brasil. Cartilha do Censo 2010: Pessoa com Deficiência. Secretaria de Direitos Humanos da Presidência da República (SDH/PR). 2012. Secretaria Nacional de promoção dos Dire- itos da Pessoa com Deficiência (SNPD). Available at <http://www.unievangelica.edu.br/ novo/img/nucleo/cartilha-censo-2010-pessoas-com-deficienciareduzido.pdf> Accessed on Octuber 04, 2022.

Brasil. Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). Lei nº 13.146. 2015. República Federativa do Brasil, Brasília, DF. Available at <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13146.htm#:~: text=Art.%201%C2%BA%20%C3%89%20institu%C3%ADda%20a,Par%C3%A1grafo% 20%C3%BAnico.> Accessed on Octuber 04, 2022.

BRASIL. Lei Nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD), Redação dada pela Lei nº 13.853, de 2019 (in Portuguese). 2019. República Federativa do Brasil, Brasília, DF. Available at <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm>. Accessed on June 26, 2020.

BRASIL. Rede Nacional de Dados em Saúde - RNDS. 2020. Ministério da Saúde. Available at <https://www.gov.br/saude/pt-br/composicao/seidigi/rnds>. Accessed on July 18, 2023.

Brasil. Cadastro Inclusão da Pessoa com Deficiência. 2022. Available at<https://www.gov.br/mdh/pt-br/navegue-por-temas/pessoa-com-deficiencia/ acoes-e-programas/cadastro-inclusao-da-pessoas-com-deficiencia>. Accessed on Novem- ber 7, 2022.

Brasil. Cadastro Inclusão facilita acesso de pessoas com deficiência a políticas públicas. 2022. Available at <https://www.gov.br/pt-br/noticias/assistencia-social/2022/

07/cadastro-inclusao-facilita-acesso-de-pessoas-com-deficiencia-a-politicas-publicas>.Accessed on November 10, 2022.

BYABAZAIRE, J.; O’HARE, G.; DELANEY, D. Data quality and trust: Review of challenges and opportunities for data sharing in iot. Electronics, v. 9, n. 12, 2020. Disponível em: <https://www.mdpi.com/2079-9292/9/12/2083>.

CAI, L.; ZHU, Y. The challenges of data quality and data quality assessment in the big data era. Data Science Journal, v. 14, n. 2, May 2015.

CAMPOS, R. T. O. et al. Oficinas de construção de indicadores e dispositivos de avaliação: uma nova técnica de consenso. Estudos e pesquisas em psicologia, Universidade do Estado do Rio de Janeiro, v. 10, n. 1, p. 221–241, 2010.

CHEN, M. et al. Survey on data quality. In: 2012 World Congress on Information and Communication Technologies. [S.l.: s.n.], 2012. p. 1009–1013.

CMMI Institute. Data Management Maturity (DMM) Model. [S.l.], 2014. 248 p.

COUNCIL, E. D. M. Data Management Capability Assessment Model (DCAM). [S.l.], 2014. 53 p.

CUANI, A. L. Conheça os 8 principais tipos de Transtornos Mentais e sua prevalência. 2021. Disponível em: <https://zenklub.com.br/blog/palavra-de-especialista/conheca-os-8-principais-tipos-de-transtornos-mentais-e-sua-prevalencia/>. Acesso em: 03 de novembro de 2022.

DONG, S.; LUCAS, M. S. Psychological profile of university students with different types of disabilities. Journal of College Student Development, v. 55, n. 5, p. 481–485, July 2014.

DONG, S.; LUCAS, M. S. An analysis of disability, academic performance, and seeking support in one university setting. Career Development and Transition for Exceptional Individuals, v. 39, n. 1, p. 47–56, 2016.

DUARTE, E. S. R. Sintomas comportamentais e psicológicos na demência no contexto da longa permanência: avaliação de idosos e intervenção para cuidadores. 144 p. Tese (D. Sc. Thesis) — Universidade Estadual Paulista “Júlio de Mesquita Filho”, São Paulo, Brazil, August 2022.

DUARTE, H. ESTUDO SOBRE O DESEMPENHO DOS ESTUDANTES COM DEFICIÊNCIA NO ENEM 2019. 63 p. Dissertação (M. Sc. Thesis) — Universidade Federal de Viçosa (UFV), Minas Gerais, Brasil, 10 2020.

ECKSTEIN, D. Meaningful jobs for students with disability: from luck to business as usual. National Centre for Student Equity in Higher Education, 2022.

EPPLER, M. Managing Information Quality. Increasing the Value of Information in Knowledge-Intensive Products and Processes. [S.l.: s.n.], 2009. ISBN 978-3-540-32225-2.

European Parliament and Council of the European Union. General Data Protection Regulation. 2016. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>. Accessed on June 10, 2020.

FANTONELLI, M. et al. Lei geral de proteção de dados e a interoperabilidade na saúde pública. Journal of Health Informatics, v. 12, mar. 2021.

FILL, H.; SINDEREN, M. van; MACIASZEK, L. Software Technologies: 16th International Conference, ICSOFT 2021, Virtual Event, July 6–8, 2021, Revised Selected Papers. Springer International Publishing, 2022. (Communications in Computer and Information Science). ISBN 9783031115134. Disponível em: <https://books.google.com.br/books?id= c4d8EAAAQBAJ>.

FRAUSIN, M. Empathy in the design process: A workshop experience on temporary and situational disabilities with design students. In: EDULEARN22 Proceedings. IATED, 2022. (14th International Conference on Education and New Learning Technologies), p. 8076–8082. ISBN 978-84-09-42484-9. ISSN 2340-1117. Disponível em: <https://dx.doi. org/10.21125/edulearn.2022.1914>.

FREITAS, I. R. d. et al. Liderança em enfermagem no contexto hospitalar: percepção de enfermeiros gestores. Revista Recien - Revista Científica de Enfermagem, v. 12, n. 40, p. 93–102, dez. 2022.

GANGEMI, A.; TENORE, K.; MANCINI, F. Two reasoning strategies in patients with psychological illnesses. Frontiers in Psychology, v. 10, p. 2335, 10 2019.

GATTE, C. C. O transtorno bipolar na infÂncia e seu desenvolvimento pessoal e educa- tivo. Revista eletrônica da faculdade Invest de Ciências e Tecnologia, v. 4, n. 2, Octuber 2021.

GOOGLE. Upcoming Updates to the Google Ads Data Protection Terms. 2020. Available at <https://privacy.google.com/businesses/lgpdtermsupdates/?hl=en>. Accessed on July 26, 2020.

Governo do Estado do Parana. Deficiência Física. 2022. Disponível em :<https://www.justica.pr.gov.br/Pagina/DeficienciaFisica#:~:text=S\%C3\%A3o\%20altera\%C3\%A7\%C3\%B5es\%20completas\%20ou\%20parciais,a\%20fala%2C\%20em\%20diferentes\%20graus>. Acesso em: 01 de novembro de 2022.

Governo do Estado do Rio de Janeiro. CARTEIRA DIFERENCIADA E CARTÃO PcD. 2023. Disponível em: <https://www.detran.rj.gov.br/_documento.asp?cod=10844>. Acesso em: 12 de fevereiro de 2023.

GRILLO, E.; SILVA, R. J. M. d. Manifestações precoces dos transtornos do comporta- mento na criança e no adolescente. Jornal de Pediatria, Sociedade Brasileira de Pediatria, v. 80, p. 21–27, Apr 2004. ISSN 0021-7557. Disponível em: <https://doi.org/10.1590/ S0021-75572004000300004>.

JAYAWARDENE, V.; SADIQ, S.; INDULSKA, M. An analysis of data quality dimensions. Queensland, Australia, 2015. 32 p. First round.

JUNQUEIRA, R.; MARTINS, D.; LACERDA, C. Politica de acessibilidade e exame nacional do ensino mÉdio (enem). Revista Educação & Sociedade, v. 38, p. 453–471, 6 2017.

KUMAR, A. et al. Barriers in accessing rehabilitative services among person with disability in tertiary care hospital of northern india: A descriptive cross sectional study. Religion, v. 85, p. 83–3, 2021.

MA, F. Diagnostic and statistical manual of mental disorders-5 (dsm-5). In: Encyclopedia of Gerontology and Population Aging. [S.l.]: Springer, 2022. p. 1414–1425.

MAIA, M. I. R.; CONFORTIN, H. Tdah e aprendizagem: um desafio para a educação. Revista Perspectiva, v. 39, n. 148, p. 73–84, 2015.

MARCONI, M. d. A.; LAKATOS, E. M. Fundamentos de metodologia científica. 5. ed. [S.l.]: Atlas S.A., 2003. 303 p.

MARSH, R. Drowning in dirty data? it’s time to sink or swim: A four-stage methodology for total data quality management. Journal of Database Marketing & Customer Strategy Management, v. 12, p. 105–112, 2005.

MATTOS, L.; FARIA, F. Da creche ao ensino médio, 17% dos brasileiros estão em escolas particulares. 2022. Disponível em: <https://www1.folha.uol.com.br/educacao/2022/08/da-creche-ao-ensino-medio-so17-dos-brasileiros-estao-em-escolas-particulares>. Acesso em: 22 de março de 2023.

MCGILVRAY, D. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted InformationTM. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008. ISBN 9780080558394.

MOURÃO, A. B.; NETTO, J. F. de M. Simroaa multi-agent recommendation system for recommending accessible learning objects. 2019 IEEE Frontiers in Education Conference (FIE), p. 1–9, 2019.

MUNIZ, E. M.; BATISTA, M. A. A importância do serviço social na socialização do cadastro único: a experiência do cras salinas em natal/rn. 2022.

NEUROSABER. Desafios na inclusão dos alunos com dificuldades na escola. 2021. Disponível em: <https://institutoneurosaber.com.br/desafios-na-inclusao-dos-alunos-comdeficiencia-na-escola/>. Acesso em: November 7, 2022.

NICHOLSON, N. et al. Dotting the “i” of interoperability in fair cancer-registry data sets. In: KAIS, G.; HAMDI, Y. (Ed.). Cancer Bioinformatics. Rijeka: IntechOpen, 2021. cap. 7. Disponível em: <https://doi.org/10.5772/intechopen.101330>.

NOGUERA, V.; BRANCO, K. C.; CIFERRI, C. Análise de desempenho das mulheres no enem. Brazilian Journal of Development, v. 6, p. 35716–35737, 01 2020.

OLIVEIRA, B. R. de. DIFICULDADES ENCONTRADAS PELOS DA EDUCAÇÃO INFANTIL NO PROCESSO DE INCLUSÃO. 20 p. Undergraduate (B.Sc.) Final Project —Instituto Federal Goiano, Goiás, Brazil, February 2022.

PCD, J. D. Número de alunos da educação especial na rede regular de ensino aumenta quase seis vezes em 23 anos. 2023. Disponível em: <https://diariopcd.com.br/2023/03/19/numero-de-alunos-da-educacao-especial-na-rede-regular-de-ensino-aumenta-quase-seis-vezes-em-23->.Acesso em: 21 de abril de 2023.

PERRONE, C.; STRASSBURGE, S. Privacy and Data Protection – From Europe to Brazil. Panorama of Brazilian Law, v. 6, 2018.

PICHLER, R. F. User-capacity toolkit: conjunto de ferramentas para guiar equipes mul- tidisciplinares nas etapas de levantamento, organização e análise de dados em projetos de tecnologia assistiva. 310 p. Tese (D. Sc. Thesis) — Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2019.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. Communications of the ACM, v. 45, p. 211—218, April 2002.

RANJIT, S.; KAWALJEET, S. A descriptive classification of causes of data quality prob- lems in data warehousing. International Journal of Computer Science Issues, v. 7, 05 2010.

REDMAN, T. C. The impact of poor data quality on the typical enterprise. Communications of the ACM, v. 41, February 1998.

REDMAN, T. C. Data Quality: The Field Guide. [S.l.]: Digital Press, 2001. 260 p.

REDMAN, T. C. Data: An unfolding quality disaster. DM Review, August 2004.

REDMAN, T. C.; GODFREY, A. B. Data Quality for the Information Age. 1st. ed. USA: Artech House, Inc., 1997. ISBN 0890068836.

REIS, F. d. Modelagem de Dados -– Restrições de Integridade em Bancos de Dados. 2020. Disponível em: <http://www.bosontreinamentos.com.br/modelagem-de-dados/ modelagem-de-dados-restricoes-de-integridade/>. Acesso em: 6 de Janeiro.

RESENDE, M. S. d.; PONTES, S.; CALAZANS, R. O dsm-5 e suas implicações no processo de medicalização da existência. Psicologia em Revista, scielopepsic, v. 21, p. 534 – 546, 09 2015. ISSN 1677-1168. Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677116820150003000008&nrm=iso>.

SANTOS, W. A. d. Cadastro inclusão : a construção de um instrumento de gestão de políticas públicas para pessoas com deficiência. 40 p. Undergraduate (B.Sc.) Final Project —Universidade de Brasília (UnB), Brasília, Brazil, November 2018.

SANUSI, R. A. et al. Transitions between versions of the international classification of diseases and chronic disease prevalence estimates from administrative health data:

a population-based study. In: BMC Public Health. Manitoba, Canada: [s.n.], 2022. (BMC), p. 1–11.

SCANNAPIECA, M.; BATINI, C. Data Quality: Concepts, Methodologies and Techniques. [S.l.]: Springer Berlin Heidelberg, 2006. 262 p.

SCHEMITSCH, C.; NAUTH, A. Psychological factors and recovery from trauma. Injury, v.51, p. S64–S66, 2020. ISSN 0020-1383. Optimizing Patient Function After Musculoskeletal Trauma. Disponível em: <https://www.sciencedirect.com/science/article/pii/ S0020138319306886>.

SERHANI, M. et al. An hybrid approach to quality evaluation across big data value chain. In: . [S.l.: s.n.], 2016.

SIDI, F. et al. Data quality: A survey of data quality dimensions. In: 2012 International Conference on Information Retrieval Knowledge Management. [S.l.: s.n.], 2012. p. 300– 304.

SIGNAL. Introduction to Unified Data. 2022. Disponível em: <https://mixpanel.com/ blog/introduction-unified-data/#:~:text=Unified>. Acesso em: 16 de maio de 2023.

SILVA, M. D. A. D. Direito, era digital e inclusÃo: o acesso à educação para pessoa com deficiência. In: III Congresso Nacional de Ciência e Educação. Rio Grande do Norte, Brasil: [s.n.], 2022. p. 40–43.

SILVEIRA, F. L.; BARBOSA, M.; SILVA, R. Exame nacional do ensino médio (enem): Uma análise crítica. Revista Brasileira de Ensino de Física, v. 37, 3 2015.

SMARTBRIDGE. Data Done Right: 6 Dimensions of Data Quality. 2020. Disponível em:<https://smartbridge.com/data-done-right-6-dimensions-of-data-quality/>. Acesso em: 07 de outubro de 2022.

SORRENTINO, T. A. @dapt - UM MODELO SEMÂNTICO PARA PROVER ACESSIBILIDADE PARA DEFICIENTES VISUAIS NA INTERNET. 72 p. Dissertação

(Dissertação de Mestrado) — Universidade do Estado do Rio Grande do Norte e Universidade Federal Rural do Semi-Árido, Mossoró, Rio Grande do Norte, Brasil, 2013.

TANNO, L. K. et al. Global implementation of the world health organization's international classification of diseases (icd)-11: The allergic and hypersensitivity conditions model. Allergy, Wiley Online Library, v. 75, n. 9, p. 2206–2218, 2020.

TERRA, R. N.; GOMES, C. G. Inclusão escolar: carências e desafios da formação e atu- ação profissional. Revista Educação Especial, v. 26, n. 45, p. 109–123, dez. 2012. Disponível em: <https://periodicos.ufsm.br/educacaoespecial/article/view/5629>.

TRAN, H. X. et al. Intervention recommendation for improving disability employment. In: IEEE. 2020 IEEE International Conference on Big Data (Big Data). [S.l.], 2020. p. 1671–1680.

VALLE, A. E. d. O. d. Transtorno do Déficit de Atenção e Hiperatividade. 17 p. Under- graduate (B.Sc.) Final Project — Centro Universitário Internacional Uninter, Curitiba, Brasil/Paraná, May 2012.

VIEIRA, C. Quem você exclui no seu programa de inclusão? 2022. Disponível em: <https://www.linkedin.com/pulse/quem-voc\%C3\%AA-exclui-seu-programa-de-inclus\%C3\%A3o-candida-vieira/?trk=pulse-article&originalSubdomain=pt>. Acesso em: 02 de outubro de 2022.

VIGGIANO, E.; MATTOS, C. O desempenho de estudantes no enem 2010 em diferentes regiões brasileiras. Revista Brasileira de Estudos Pedagógicos, v. 94, 06 2019.

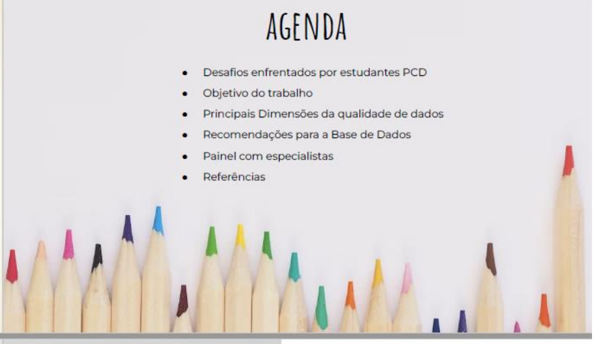
WANG, R. Y.; STRONG, D. M. Beyond accuracy: What data quality means to data consumers. Journal of Management Information System, v. 12, p. 5–33, 1996.

WU, Y. et al. A fast projection-based algorithm for clustering big data. Interdisciplinary Sciences: Computational Life Sciences, v. 11, n. 3, p. 360–366, June 2018.

ZHANG, D. Big data security and privacy protection. In: 8th International Conference on Management and Computer Science (ICMCS 2018). Shenyang, China: [s.n.], 2018. p. 275–278.

APPENDIX A - Slide presentation for the technical assessment

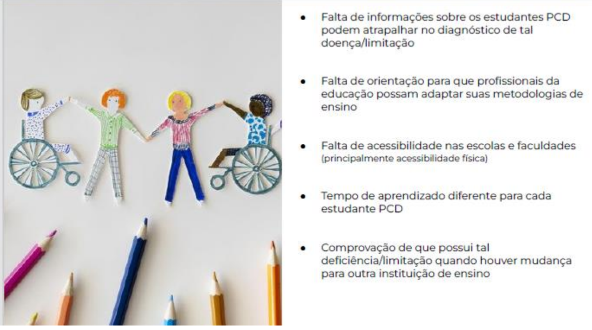
EDUCAÇÃO INCLUSIVA



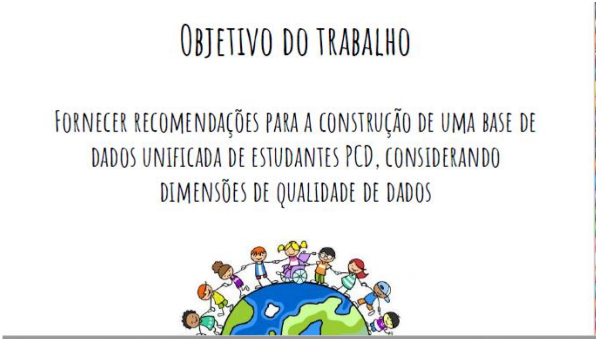
AGENDA

- Desafios enfrentados por estudantes PCD
- Objetivo do trabalho
- Principais Dimensões da qualidade de dados
- Recomendações para a Base de Dados
- Painel com especialistas
- Referências

DESAFIOS ENFRENTADOS POR ESTUDANTES PCD



- Falta de informações sobre os estudantes PCD podem atrapalhar no diagnóstico de tal doença/limitação
- Falta de orientação para que profissionais da educação possam adaptar suas metodologias de ensino
- Falta de acessibilidade nas escolas e faculdades (principalmente acessibilidade física)
- Tempo de aprendizado diferente para cada estudante PCD
- Comprovação de que possui tal deficiência/limitação quando houver mudança para outra instituição de ensino

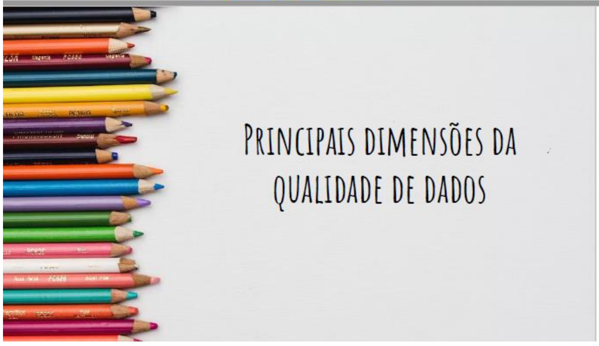


OBJETIVO DO TRABALHO

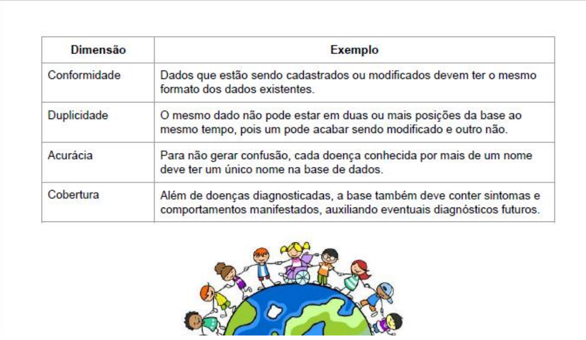
FORNECER RECOMENDAÇÕES PARA A CONSTRUÇÃO DE UMA BASE DE DADOS UNIFICADA DE ESTUDANTES PCD, CONSIDERANDO DIMENSÕES DE QUALIDADE DE DADOS

PRINCIPAIS DIMENSÕES DA QUALIDADE DE DADOS

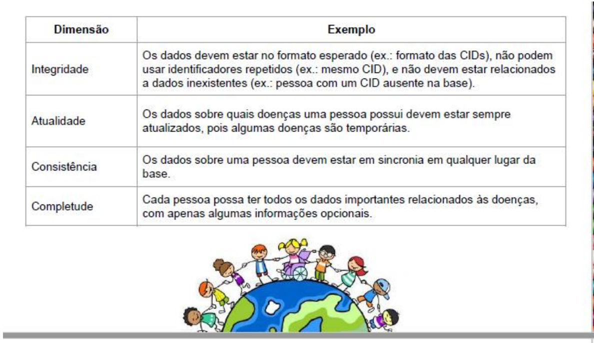
Integridade	Conformidade
Atualidade	Duplicidade
Consistência	Acurácia
Completude	Cobertura



PRINCIPAIS DIMENSÕES DA QUALIDADE DE DADOS



Dimensão	Exemplo
Conformidade	Dados que estão sendo cadastrados ou modificados devem ter o mesmo formato dos dados existentes.
Duplicidade	O mesmo dado não pode estar em duas ou mais posições da base ao mesmo tempo, pois um pode acabar sendo modificado e outro não.
Acurácia	Para não gerar confusão, cada doença conhecida por mais de um nome deve ter um único nome na base de dados.
Cobertura	Além de doenças diagnosticadas, a base também deve conter sintomas e comportamentos manifestados, auxiliando eventuais diagnósticos futuros.



Dimensão	Exemplo
Integridade	Os dados devem estar no formato esperado (ex.: formato das CIDs), não podem usar identificadores repetidos (ex.: mesmo CID), e não devem estar relacionados a dados inexistentes (ex.: pessoa com um CID ausente na base).
Atualidade	Os dados sobre quais doenças uma pessoa possui devem estar sempre atualizados, pois algumas doenças são temporárias.
Consistência	Os dados sobre uma pessoa devem estar em sincronia em qualquer lugar da base.
Completude	Cada pessoa possa ter todos os dados importantes relacionados às doenças, com apenas algumas informações opcionais.

Recomendações para a Base de Dados

Dados Pessoais

- Nome
- CPF (chave primária)
- Telefone
- Email
- Data de Nascimento
- Gênero
- Endereço - País
- Endereço - Estado
- Endereço - Cidade
- Endereço - Bairro

Conformidade
Integridade
Duplicidade
Consistência
Completude

Dados relacionados à doença

- Doença Temporária ou Permanente
- Data de início do diagnóstico
- Data final da doença (se existir)
- Indicar se a doença permanece presente no momento
- Indicar se o diagnóstico é oficial ou não

Integridade
Atualidade
Cobertura
Completude

Dados relacionados ao Histórico

- Criar um histórico comportamental dos estudantes ao longo da sua vida escolar e acadêmica.
- Exemplo: Criar uma tabela de logs, onde cada registro teria a data e o comportamento observado do estudante.

Atualidade
Cobertura
Completude

Dados relacionados à classificação

- Utilizar a Classificação Internacional de Doenças (CID)
- Agrupar doenças conforme a natureza da deficiência:
  - Distúrbios Cognitivos
  - Psicológicos
  - Físicos

Cobertura
Conformidade
Integridade
Duplicidade
Consistência
Acurácia

Atualização dos Dados

- Ter um mecanismo de confirmação periódica das informações dos estudantes.
- Em doenças permanentes novos sintomas e limitações que prejudiquem o desempenho escolar de estudantes PCD podem surgir ao longo do tempo.
- A base de dados precisa ser flexível a mudanças.

Integridade
Atualidade
Consistência

Integridade dos Dados

- **Integridade de domínio:** campos de preenchimento obrigatório e valores pré-determinados.
- **Integridade referencial:** uso de chaves primárias e estrangeiras para fazer o relacionamento entre diferentes tabelas da base.
- **Integridade do vazio:** definir quais campos são obrigatórios ou não, para evitar valores nulos onde não poderiam existir.
- **Integridade de chave:** garantir que os valores dos campos de chave primária sejam sempre únicos e estejam preenchidos.

Integridade
Consistência

DEMAIS RECOMENDAÇÕES

- Utilizar **campos identificadores únicos**. Para todos os campos que já possuem um valor pré-determinado, é possível criar dimensões para esses valores e atribuir IDs para que a associação dos dados se dê por meio de chaves.
- Determinados campos devem ter **preenchimento obrigatório** (exemplos: cpf, name, email, flag_active, oficial_diagnosis).
- Criar uma **dimensão** com as informações de CID e relacionar elas através de chaves.
- Trazer **informações já existentes** de bases do governo, como dados do Detran, por exemplo.

PAINEL COM ESPECIALISTAS

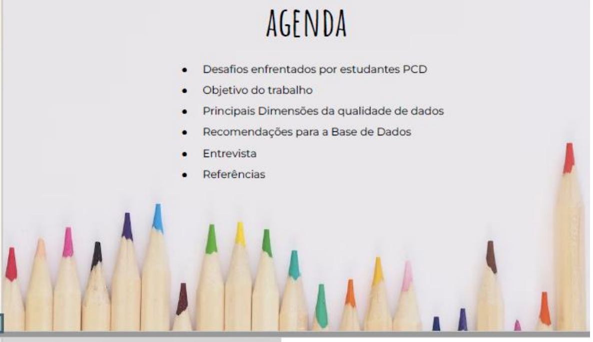
REFERÊNCIAS

- **As imagens da apresentação foram retiradas dos sites:**
  - https://unsplash.com
  - https://www.unesc.net/portal/blog/ver/53/38136
  - https://blog.psicologiaviva.com.br/inclusao-social/
  - https://sesisc.org.br/pt-br/para-voce/educacao-inclusiva
  - https://revistaensinosuperior.com.br/questao-de-inclusao/
  - https://uvagpclass.wordpress.com/2017/11/20/a-tecnica-do-brainstorming/

OBRIGADA!

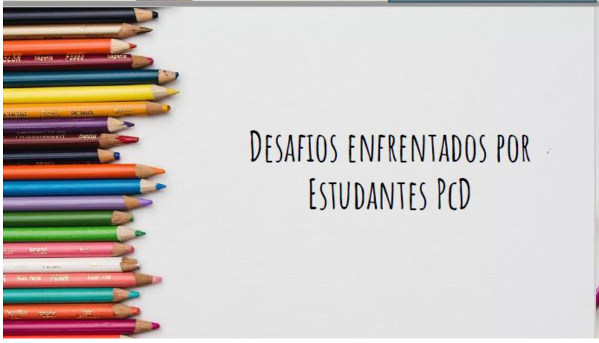
APPENDIX B - Slide presentation for the educational assessment

Educação Inclusiva

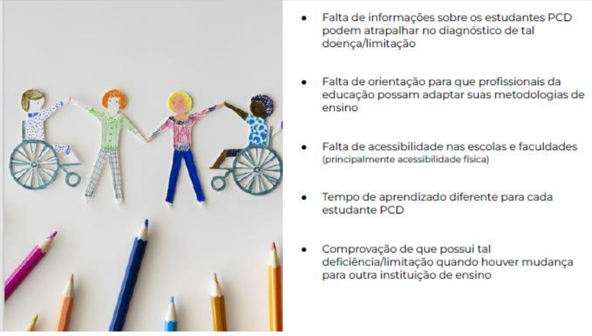


AGENDA

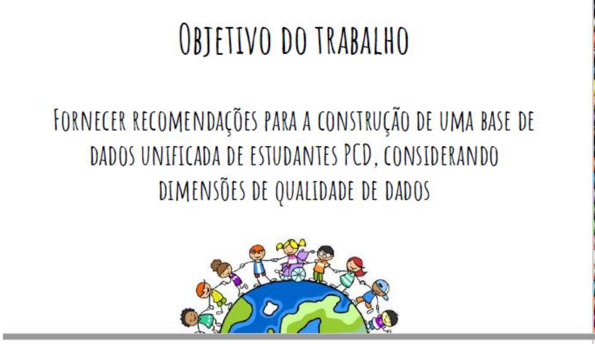
- Desafios enfrentados por estudantes PCD
- Objetivo do trabalho
- Principais Dimensões da qualidade de dados
- Recomendações para a Base de Dados
- Entrevista
- Referências



Desafios enfrentados por Estudantes PcD

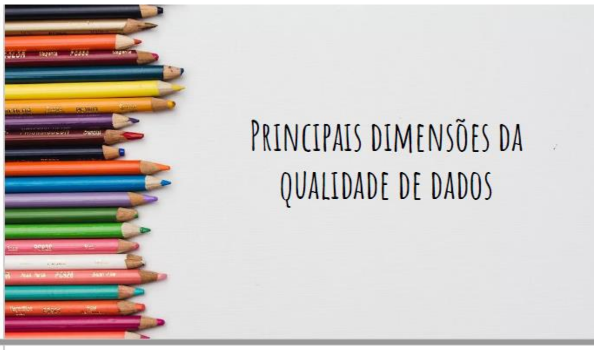


- Falta de informações sobre os estudantes PCD podem atrapalhar no diagnóstico de tal doença/limitação
- Falta de orientação para que profissionais da educação possam adaptar suas metodologias de ensino
- Falta de acessibilidade nas escolas e faculdades (principalmente acessibilidade física)
- Tempo de aprendizado diferente para cada estudante PCD
- Comprovação de que possui tal deficiência/limitação quando houver mudança para outra instituição de ensino

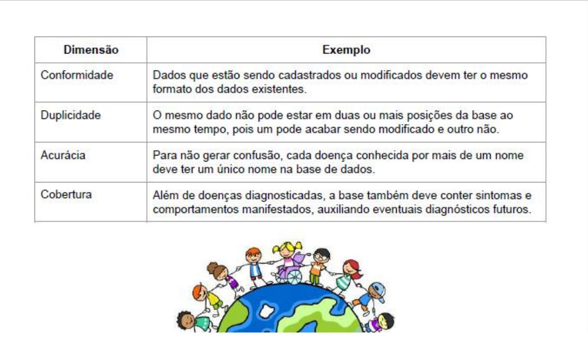


Objetivo do trabalho

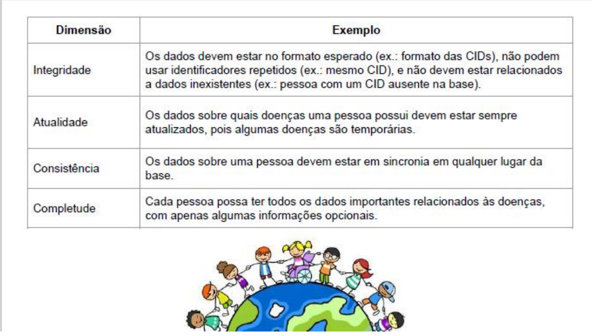
Fornecer recomendações para a construção de uma base de dados unificada de estudantes PCD, considerando dimensões de qualidade de dados



Principais dimensões da qualidade de dados



Dimensão	Exemplo
Conformidade	Dados que estão sendo cadastrados ou modificados devem ter o mesmo formato dos dados existentes.
Duplicidade	O mesmo dado não pode estar em duas ou mais posições da base ao mesmo tempo, pois um pode acabar sendo modificado e outro não.
Acurácia	Para não gerar confusão, cada doença conhecida por mais de um nome deve ter um único nome na base de dados.
Cobertura	Além de doenças diagnosticadas, a base também deve conter sintomas e comportamentos manifestados, auxiliando eventuais diagnósticos futuros.



Dimensão	Exemplo
Integridade	Os dados devem estar no formato esperado (ex.: formato das CIDs), não podem usar identificadores repetidos (ex.: mesmo CID), e não devem estar relacionados a dados inexistentes (ex.: pessoa com um CID ausente na base).
Atualidade	Os dados sobre quais doenças uma pessoa possui devem estar sempre atualizados, pois algumas doenças são temporárias.
Consistência	Os dados sobre uma pessoa devem estar em sincronia em qualquer lugar da base.
Completude	Cada pessoa possa ter todos os dados importantes relacionados às doenças, com apenas algumas informações opcionais.

RECOMENDAÇÕES PARA A BASE DE DADOS

DADOS PESSOAIS

- Nome
- CPF (chave primária)
- Telefone
- Email
- Data de Nascimento
- Gênero
- Endereço - País
- Endereço - Estado
- Endereço - Cidade
- Endereço - Bairro

Conformidade
Integridade
Duplicidade
Consistência
Completude

DADOS RELACIONADOS À DOENÇA

- Doença Temporária ou Permanente
- Data de início do diagnóstico
- Data final da doença (se existir)
- Indicar se a doença permanece presente no momento
- Indicar se o diagnóstico é oficial ou não

Integridade
Atualidade
Cobertura
Completude

DADOS RELACIONADOS AO HISTÓRICO

- Criar um histórico comportamental dos estudantes ao longo da sua vida escolar e acadêmica.
- Exemplo: Criar uma tabela de logs, onde cada registro teria a data e o comportamento observado do estudante.

Atualidade
Cobertura
Completude

DADOS RELACIONADOS À CLASSIFICAÇÃO

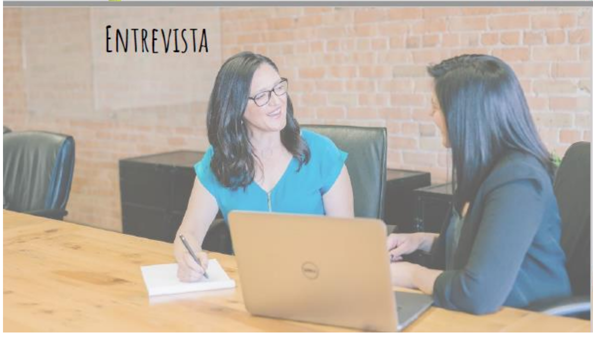
- Utilizar a Classificação Internacional de Doenças (CID)
- Agrupar doenças conforme a natureza da deficiência:
  - Distúrbios Cognitivos
  - Psicológicos
  - Físicos

Cobertura
Conformidade
Integridade
Duplicidade
Consistência
Acurácia

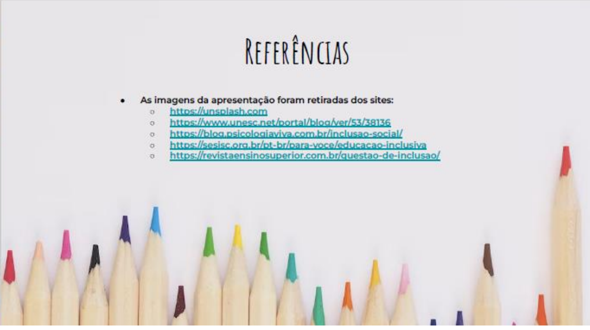
ATUALIZAÇÃO DOS DADOS

- Ter um mecanismo de confirmação periódica das informações dos estudantes.
- Em doenças permanentes, novos sintomas e limitações que prejudicam o desempenho escolar de estudantes PCD podem surgir ao longo do tempo.
- A base de dados precisa ser flexível a mudanças.

Integridade
Atualidade
Consistência



ENTREVISTA



REFERÊNCIAS

- As imagens da apresentação foram retiradas dos sites:
  - https://unsplash.com
  - https://www.unesc.net/portal/blog/ver/53/38136
  - https://blog.psicologiaviva.com.br/inclusao-social/
  - https://sesisc.org.br/pt-br/para-voce/educacao-inclusiva
  - https://revistaensinosuperior.com.br/questao-de-inclusao/

OBRIGADA!