



The Sleep Revolution Platform: a Dynamic Data Source Pipeline and Digital Platform Architecture for Complex Sleep Data

Bjarki Freyr Sveinbjarnarson^{1,2} · Lisa Schmitz^{1,2} · Erna Sif Arnardottir² · Anna Sigridur Islind^{1,2}

Accepted: 9 March 2023 / Published online: 10 April 2023
© The Author(s) 2023

Abstract

Purpose of Review The complexity of the data collected for sleep research is increasing, and the focal point of sleep research is dependent on a higher number of data sources. Data collected for sleep studies often includes both subjective and objective measurements of sleep quality and is gathered over a more extended period, e.g., for weeks, months, or even years. However, this variety and volume of data make it challenging and time-consuming for researchers to utilize. Therefore, sophisticated data structures are necessary to utilize data in sleep research.

Recent Findings This paper explores how heterogeneous data sources can be represented in a homogeneous database design. The following research questions drove our work: (i) How can we represent sleep data from heterogeneous sources in a homogenous digital platform database? and (ii) How can a data source pipeline transform various data sources into a homogeneous data format?

Summary This paper's main contributions are conceptualizing the design and development of a homogeneous database and digital platform architecture and a data source pipeline that fits well for sleep research in particular and healthcare research in general.

Keywords Sleep Revolution · Information Systems · Digital Platform Architecture · Database Design · Data Source Pipeline

Introduction

Sleep research and medicine involve data collection from a variety of sources [1••]. The gold standard for measuring and assessing more complex sleep disorders such as sleep-disordered breathing is polysomnography (PSG), a method that traditionally involves collecting processed signals from connected sensors [2, 3]. Driven by a growing need for longitudinal sleep assessment, both for research, clinical assessment, and treatment response, additional methods to collect subject and objective sleep data have been developed such as digital symptom trackers and wearables as well as treatment compliance assessments [4]. Moreover, signal formats

generally differ between devices and manufacturers, contributing further to the data's variety.

Researching and improving longitudinal sleep assessments is one of the goals of the Sleep Revolution [1••]. The project has multiple cornerstones in different research fields, with the aim of collecting data in multiple prospective studies and processing retrospective data from thousands of completed sleep studies from different sources. Due to the quantity and complexity of the data sources, the collected data must be translated into a homogeneous data format to ensure that researchers can work efficiently with the data. Therefore, it is vital to design and develop a digital platform with a novel database architecture to store, process, and combine the various data sources for the Sleep Revolution [4].

The growing complexity of modern digital platforms due to an increase in data variety and the resulting challenges are not unique to sleep research and medicine [5•, 6]. To address these challenges, researchers have focused on advancing digital platform design by providing architecture guidelines on an abstract level, such as the layered modular architecture [7•]. While such high-level concepts

✉ Bjarki Freyr Sveinbjarnarson
bjarkis@ru.is

¹ Department of Computer Science, Reykjavik University, Reykjavik, Iceland

² Reykjavik University, Sleep Institute, Reykjavik, Iceland

enable digital platform designers to organize the complete digital platform ecosystem, De Reuver et al. emphasize the need to expand the research on digital platform architecture on different levels [5•]. In this paper, we contribute to this call by focusing on the processing and organization of data found at the data level of digital platform architectures. We present a homogeneous digital platform and database design to represent the heterogeneous data sources and a data source pipeline design for processing the data sources into a homogeneous data format. Our contributions result from a detailed analysis and categorization of the data. The following research questions drove our work: (i) How can we represent sleep data from heterogeneous sources in a homogenous digital platform database? and (ii) How can a data source pipeline transform various data sources into a homogeneous data format? The main contributions of this paper are through the conceptualization of the design and development of a homogeneous database and digital platform architecture and a data source pipeline to convert and process heterogeneous data sources into a validated homogenous format that fits well for sleep research in particular and healthcare research in general. In this paper, the Sleep Revolution provides an illustrative example to show the flexibility and dynamic capacity of the digital platform design. Based on our illustrative example, we argue that this particular digital platform design can be generalized and utilized in other contexts as it provides a dynamic approach for modern data.

Related Work

Designing and developing an information system that combines and adapts various data sources for multiple end-user groups is challenging. During the design and development processes, it is vital to preserve digital platform characteristics that have been identified as essential features for a successful information system design.

In their editorial paper, Constantinides et al. highlight key insights on the architecture of digital platforms [8]. They identify abundant data collection due to the rise of machine learning methodology and the increased digitization of diverse processes, such as those in healthcare. Therefore, the way digital platforms are designed to facilitate abundant data collection is a key element going forward. This phenomenon leads to new challenges for platformization and creates a need for research about digital platform architecture and data organization. Both should fulfill the criteria of being “stable and evolving” [9] in order to design a robust and long-lived digital platform ecosystem.

Yoo et al. [7•] illustrate how product innovation through digital capabilities of traditionally analog products influences the requirements of platform architecture. One of the digital

platform characteristics emphasized by this development is data homogenization. They argue that “unlike analog data, digital data originate from heterogenous sources and can be combined easily with other digital data to deliver diverse services” [7•]. Our paper follows this philosophy by applying it to the data structure and architecture of the SleepWell, the Sleep Revolution platform. Though collected through a variety of digital products and physical devices (mobile applications and sensors), we aim to bring the data together in a homogenous format. In a layered modular architecture, as introduced by Yoo et al., this feature contributes to the flexibility of a digital platform, keeping it open to the option of adding new digital elements to it, such as interfaces [7•].

The recombability of digital elements is highlighted by the research conducted by De Reuver et al. [5•]. To ensure that multiple applications can connect to the interface of a digital platform, data presented through that interface must not be subject to constant changes to its format. Therefore, it is advantageous if data changes and new data points can be represented within the existing data structure of the digital platform.

In recent years, digital platforms have been emerging in the healthcare industry [10]. Driven by the increasing availability of sensors and wearables for the consumer market, the wide variety of data sources drives the development of healthcare applications. However, the issue of dealing with heterogeneous data from various sources within a unified platform ecosystem arises from this situation, and research is scarce on that topic. In this paper, we contribute to that gap in the literature. Bache et al. encountered this problem when defining an architecture to combine multiple heterogeneous data sources and query them efficiently [11]. Their solution focused on the development of an abstract, reusable query model. This query model hides the underlying structure of data and enables interfaces to connect to it efficiently [11]. A new data source can be added to their architecture with a lightweight adapter. Nevertheless, this adapter must be independently developed for each new data source. The notion of lightweight adapters is similar to the seminal paper by Bygstad, which argues for the architectural vision of lightweight versus heavyweight modules [12]. However, their paper is of conceptual nature and outside of healthcare, whereas our paper contributes with digital platform architecture within healthcare specifically and with a special user case.

Methods

Action Design Research

We aimed to design and develop a digital platform that could function as a bridge between healthcare professionals, researchers, and participants and include heterogeneous data in a homogenous architecture. As we see it, it is important

to have multiple feedback loops to ensure that the digital platform design and development are aligned with the needs of the different end-user groups. Because the Sleep Revolution project size calls for a large-scale SleepWell platform, in which the abovementioned researchers are from multiple disciplines alongside the participants (that have a wide variety of needs) and the healthcare professionals, from multiple sectors, the formulation of the digital platform requirements has been a complex, iterative process. Action design research (ADR) is a method that fits well for complex and iterative research projects. ADR has four scopes: (i) problem formulation, (ii) designing solutions, (iii) reflecting upon the solutions, and (iv) learning outcomes [13]. We created an ADR workflow that builds on the four scopes for our research project (Fig. 1). Through the design and development phase, interviews were conducted continuously with different end-users.

Sleep Revolution

The Sleep Revolution is a European Union Horizon 2020 project across multiple countries and different beneficiaries. Moreover, it is a multi-disciplinary consortium with a cornerstone in multiple fields, such as sleep medicine and research, computer science, biomedical science, psychology, engineering, and sports science. One of the major objectives of the Sleep Revolution is to transform the current diagnostic methods and treatment follow-up for sleep-disordered breathing [1••]. This objective utilizes the retrospective sleep study data pool of tens of thousands of sleep recordings and health information [1••]. Another major objective is to promote participatory healthcare with technological solutions, where it aims to design a digital platform to promote participatory healthcare used

in numerous prospective studies [1••]. Coupled with that, an important step is to centralize a wide variety of sleep data from thousands of patients and research participants into unified data sets that can be accessed digitally through a novel digital platform [1••].

Data Sources

Due to the project’s aforementioned magnitude, the data comes in a variety of different formats and has been collected with different devices; ergo, the data is heterogeneous. Therefore, unifying the heterogeneous data sources into a homogenous format while also representing the data feasibly in our digital platform is a significant challenge. Before designing and developing a digital platform architecture that combines multiple sleep data sources, we mapped out key data sources collected throughout our different Sleep Revolution projects.

Sleep Studies

During an overnight sleep study, a variety of signals are collected using sensors to measure changes in physiological states that occur while a person sleeps. The study includes channels to measure electroencephalography for brain wave activity, electrooculography to measure eye movements, and chin electromyography for muscle tone. Together, these channels allow for the assessment of different sleep stages and wake periods measured in 30 s epochs throughout the night and arousals from sleep. The sleep studies also include respiratory flow assessment to assess breathing, as well as respiratory movements via thorax and abdomen belts and blood oxygen levels and pulse via a pulse oximeter. Together, these measurements allow for

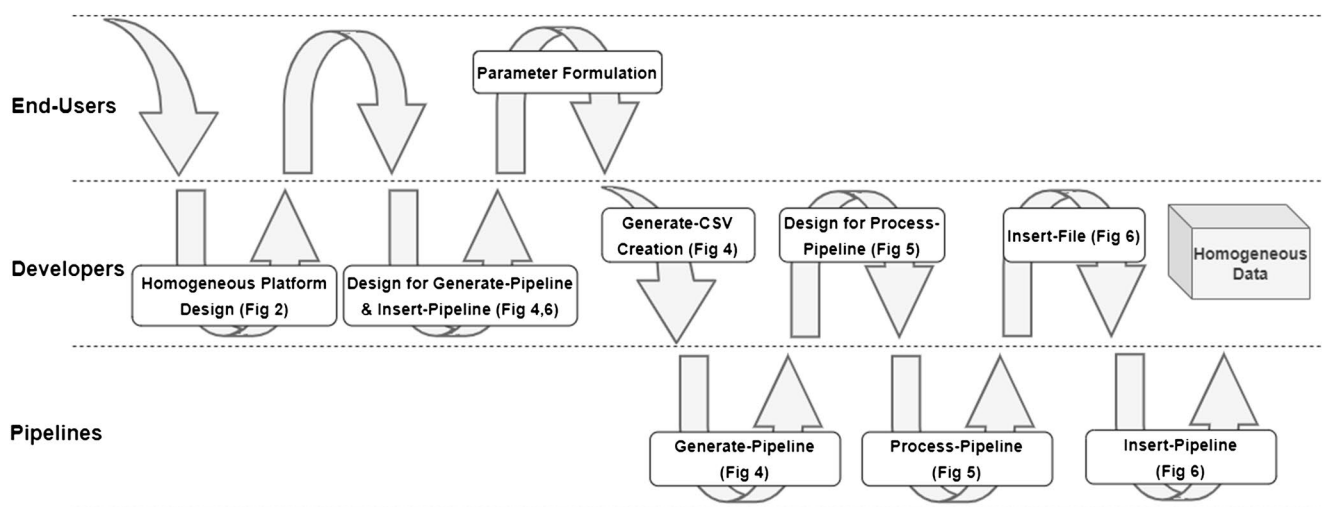


Fig. 1 An illustration of our action design research workflow for the designing of the digital platform design and the pipelines

the assessment of sleep apnea severity. Additionally, an electrocardiography, leg electromyography (for periodic leg movement assessment), body position and activity, and possibly synchronized video and audio (e.g., for snore measurement) are included. Therefore, during a sleep study, multiple sensors and devices are used to capture sleep objectively. The prospective data collection, processing, and device formatting are done in Sleep Revolution through Noxturnal (Nox Medical, Reykjavik, Iceland). Furthermore, the sleep studies are manually scored using Noxturnal as well. The results from the scored PSG are split into 16 different categories, e.g., “position activity” or “respiratory.” The 16 categories represent over 1700 unique classifications of data. Examples of classification are “SleepTotalN3Duration,” the sleep duration in the N3 sleep stage, and “SnoringTrainsPerHourSupine,” the number of “snoring trains” per hour in a supine sleeping position. The high number of unique classifications is due to the high number of scored events, like apnea, hypopnea, snoring, and arousal, which can be further distinguished by, e.g., the sleep stage and the sleeping position. Therefore, it adds up to a multitude of data. The sleep studies are exported and include (i) raw signal files using a European Data Format (EDF) and (ii) parameter files in a semi-structured Extensible Markup Language (XML) format.

Wearables

The digital platform allows wearable solutions to be connected to it. One of the most used wearables in the Sleep Revolution research is the smartwatch Scanwatch (Withings, Paris, France). The Scanwatch has a photoplethysmography sensor and a 3-axis accelerometer, enabling it to track exercise, sleep, heart rate, and more. The smartwatch is connected to the user’s phone via Bluetooth and gathers data from all connected Withings devices into a database that is accessible via an application programming interface (API). The SleepWell platform connects to the API to retrieve data, including exercise sessions, step counts, elevation, and sleep together with raw sensor data values. While other wearables can be integrated, the Withings watch is used in all prospective studies.

Questionnaires

Research Electronic Data Capture (REDCap) [14] is a secure web application used to manage the majority of questionnaires in the Sleep Revolution, forms for staff entry as well as informed consent by research participants. This web application allows for the creation of multiple types of questions, with a wide array of answering options, e.g., checkboxes, radio buttons, time fields, numerical input, and open-ended text answers. In the Sleep

Revolution, over twenty different questionnaires have been created, sent out, and collected, with over 3000 responses from participants.

The questionnaires have two types of data sources: (i) the setup of the questionnaire, i.e., the choices of questions and types of answers, and (ii) the results of each question for each collected answer, for each participant. The exported output files from questionnaires are semi-structured comma-separated values (CSV) files. Moreover, the European Sleep Questionnaire (ESQ) is currently being digitally designed into the SleepWell platform, with access in 15 different languages [1••].

Digital Sleep Diaries

Sleep diaries are a valuable tool for gathering subjective data for providing an overview of people’s sleep quality and habits over an extended period of time [15]. The Sleep Revolution designed and developed a mobile application (an app) with an adapted version of the Consensus Sleep Diary [15] with both a morning and evening sleep diary, also in 15 different languages. The Sleep Revolution app feeds the data directly to our digital platform architecture. The sleep diary within the app is used to collect longitudinal data on subjective sleep quality and habits over a period of 3 months [16].

Cognitive Tests

The Sleep Revolution app feeds data directly into SleepWell. The app also includes a cognitive battery to measure cognitive function over an extended period of time. A cognitive battery is a collection of different cognitive tasks done in a row where each cognitive task targets specific cognitive processes or domains, i.e., perceptual skills, processing speed, episodic memory, and reasoning. The cognitive tasks and batteries are used to document current cognitive ability as well as changes in cognitive ability over time in the Sleep Revolution project.

The Sleep Revolution currently uses three sources for cognitive tests: (i) an in-lab cognitive battery, (ii) an at-home cognitive battery completed directly within the SleepWell platform by the participant, and (iii) four cognitive tests that are included in the Sleep Diary app. The data format varies for the in-lab cognitive battery since some tasks make use of photos and other types of complex input. The in-lab cognitive battery was designed in the software Inquisit (Millisecond, Seattle, USA) and outputs a semi-structured CSV file. The latter two sources send their processed results using our API directly through our digital platform, to the database.

Other Data Sources

In addition to the data sources mentioned above, there is a wide variety of other data sources collected in Sleep Revolution. We further mapped these out in Table 1.

Each data source requires a unique preprocessing approach in order to be functional, depending on how it was collected and the original format. The difficulty of preprocessing depends on its given format, if it includes manual inputs, is reliant on other data sources, and other similar factors. To provide an overview of the challenges, we have mapped them out in Table 2.

Results

First, we present the homogenous database and digital platform design and thereafter present the data source pipeline. The conceptualization of the two parts outlines the main contribution of this paper.

Homogeneous Database and Digital Platform Design

We arrived at a digital platform design that is simple to use and understand, yet flexible and dynamic enough to incorporate the various data sources. The generalized design of the database, on which the digital platform rests, is represented in Fig. 2. A five-fold design is used as the core model to

represent the data: (i) entry, (ii) form, (iii) entry-result, (iv) form-result, and (v) owner. An entry represents a data point, and a form is a collection of the entries, e.g., one entry can be one question in a form which is one questionnaire. Therefore, a data source can have multiple forms, like multiple questionnaires. The entry-result and form-result tables, on the other hand, store the individual answers and total answers to those questions and questionnaires, respectively. Therefore, every entry and form are only created once, but each entry-result and form-result can have none or multiple inputs.

One of the design strengths of this digital platform is its simplicity, since no additional tables are required to extend or add new data sources, independent of the data source. The design’s simplicity makes it easy to apply, for example, a wide variety of data analysis types to the data in the project’s next phase. Moreover, the simplicity allows for flexible and dynamic front-end development due to the consistency of the querying and the limited number of tables. This way, we can fit new data sources of heterogeneous quality into this dynamic design. Furthermore, the entry and entry-result tables make adding new data points to already existing forms effortless. That makes the design revolutionary and novel for research data collection since, for many data sources, it is common to add a new additional data point to a form retrospectively. Moreover, the flexible database design allows for holistic digital platform architecture. An additional novel feature is derived from the fact that the design allows exporting the data into a format or file that fits the different researcher end-users work environment or data analysis software, e.g., Python, R, or SPSS (Statistical Package

Table 1 Examples of the data sources and their format as collected in the Sleep Revolution

Source	Examples	Data format
Spreadsheets	Body measurements, complementary data, participant schedules, data ownerships, etc	XLSX, CSV, ODS
Audio recordings	Interviews, speech recordings, etc	Audio file formats
Video recordings	Sleep laboratory recordings	Video file formats
Mobile applications	Subjective data for chronic pain, menstrual cycle etc	JSON
Other wearable devices	Empatica smartwatch, Nukute collar, Dreem headband, etc	Multiple formats for each wearable

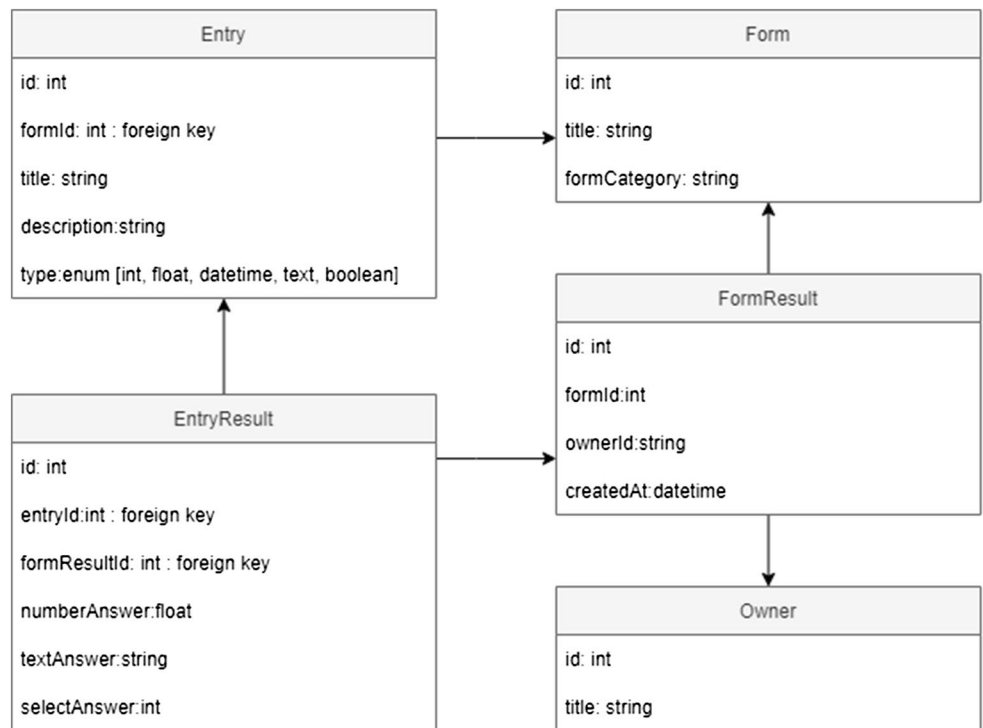
XLSX, Excel Microsoft Office Open XML Format Spreadsheet file; CSV, comma-separated values file; ODS, OpenDocument Spreadsheet; JSON, JavaScript Object Notation

Table 2 Examples of the data source categories collected in the Sleep Revolution as well as their environment’s dependencies and constrictions

Category	Uses web-app, SDK, or software	Complementary resources	API	Continuous
Questionnaires	Some	Yes	No	No
Wearables	Yes	Yes	Some	Yes
Sleep studies	Yes	Yes	No	Yes
Cognitive tests	Yes	Some	Some	No
Physical measurements	No	Some	No	No

API, application program interface; SDK, software development kit

Fig. 2 The homogeneous digital platform design fits all the different data sources. Instead of adding new tables and columns for every data source, the five-fold core model adds them as inputs, therefore reducing the size and complexity of the database considerably



for the Social Sciences). The design uses the ownership table to connect the data sources, a crucial element for research, i.e., the ability to connect all data points within a research participant vs. between research participant for analysis.

Data Source Pipeline

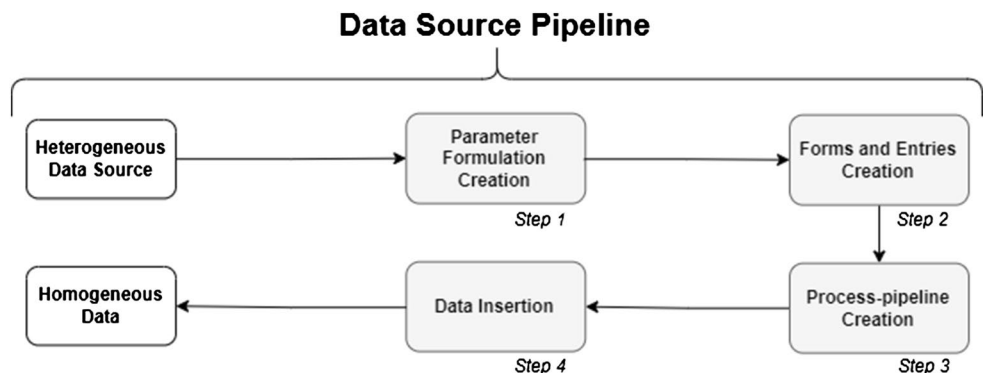
The diverse data sources need to fit into the homogeneous digital platform design, which requires processing of every data source into a homogeneous format. Since the project has a great number of data sources, we designed a generic high-level diagram for the data source pipeline that fits the data sources (Fig. 3).

The *parameter formulation creation* step is about understanding the data source and choosing the relevant data

points for end-users. The process commonly requires semi-structured interviews or collaboration between the developers, data collection team, and end-users. The selected data points are then split into one or more forms. Data sources can have multiple forms, for example, a questionnaire web application can have multiple questionnaires.

The *forms and entries creation* step consists of taking the decided parameters in the previous step and creating a comma-separated values (CSV) file for each form containing: (i) parameters, (ii) data type (integers [int], Boolean [true/false], date, etc.), and (iii) description of parameter. The step uses its own pipeline to create the form and entries (Fig. 4). The pipeline simplifies the creation of new forms and entries, only requiring a CSV containing entry names, data types, and optional descriptions. To ensure homogeneity between forms and entries, all data sources use the same

Fig. 3 The data source pipeline converts a heterogeneous data source into homogeneous data in four steps



script with a strict evaluation. The script uses the form’s title, the form’s category, and CSV as input to create the form and entries.

The *process-pipeline creation* step revolves around creating a pipeline to transform the unprocessed data source into a processed result data (Fig. 5) so that it can be automatically input (Fig. 6). The complexity of each sub-step in the process-pipeline depends on the data source environment, as stressed in the examples given in the “Data Sources” chapter. Therefore, each data source requires different approaches. In the process-pipeline, we illustrate six crucial steps to translate the data sources into a homogeneous format (Fig. 5). As shown in Table 2, some data sources rely on other software, web-apps, APIs, or SDKs for collecting the data. That makes it necessary to *export a selection of data* in the data source’s given data format. The exported data can be in multiple files or in difficult-to-use formats, requiring scripts or manual work to *combine, transform, or convert the data* into the developer’s preferred

data formats. In this research context, the data sources require ownership, that is, which participant the data result belongs to. However, the software, digital platforms, apps, and other data sources often do not contain a feature to add ownership. Therefore, it is essential to use additional resources, such as spreadsheets, to *set ownership to each data point* and to ensure data integrity. The collected research data may contain duplicates and incomplete or unusable entries, making it necessary to *clean the data*. Moreover, data sources often rely on additional resources, manual work, or manual inputs, making it essential to review and *validate the data*. After the data points have been preprocessed, cleaned, validated, and added ownership onto, it is necessary to *convert the data to input-data* that is accepted by the insert-pipeline (see Fig. 6).

The *data insertion* step is about using the insert-pipeline (Fig. 6) to transform the processed result data into form results and entry results. The insert-pipeline outlines a collection of pipelines to insert processed data results into the

Fig. 4 The generate-pipeline takes comma-separated values (CSV) as input to create forms and individual entries for data sources

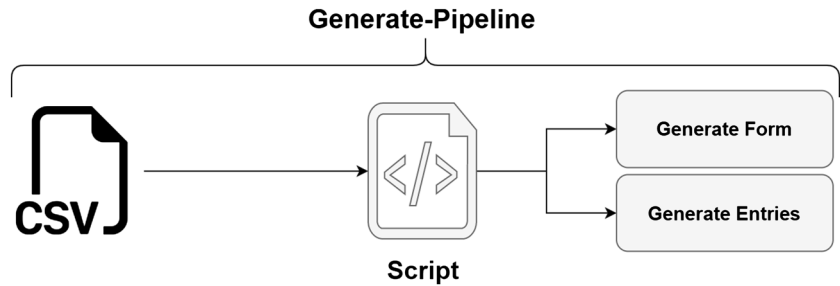


Fig. 5 The process-pipeline consists of six crucial steps to convert unprocessed data source into processed result data

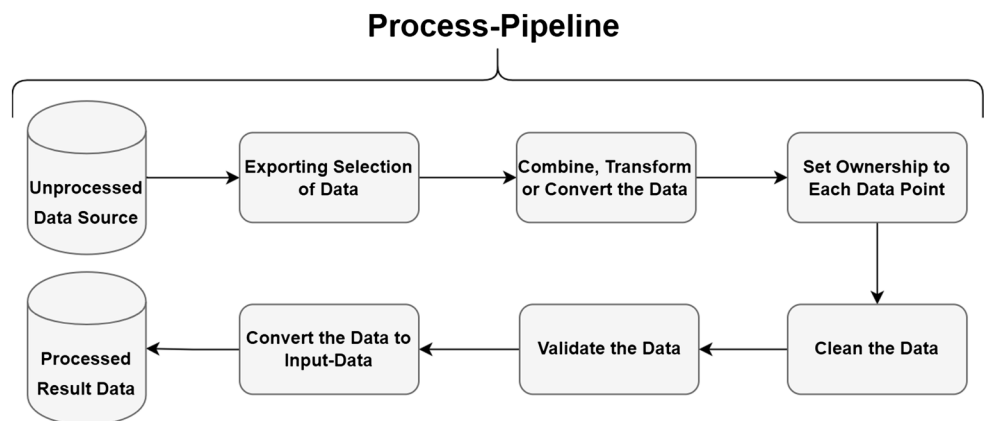
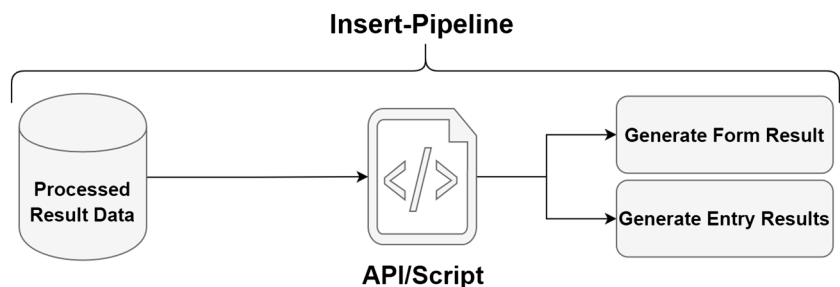


Fig. 6 The insert-pipeline converts processed result data into form results and entry results using a script or an application programming interface (API)



database, therefore connecting the data source pipeline to the homogeneous database and digital platform architecture. Some data sources consist of local files, whereas others are only accessible using an API (Table 2). In those cases, it is necessary to create its own insert-pipeline.

Discussion

In this paper, we present the design and development of a database and digital platform architecture alongside a data source pipeline to streamline data for sleep research [4]. The homogeneous database and digital platform architecture, with the addition of the data source pipeline, effectively combines heterogeneous sleep data from diverse sources into an abstract and dynamic homogeneous representation. These characteristics contribute to the digital platform's flexibility called for by Yoo et al. (2010) [7•]. Due to the adaptable data source pipeline designs, new data sources and data points can be easily added in the future by preprocessing and translating them into an accepted data format. This way, the focal point was to ensure the preservation of essential digital platform design characteristics such as reprogrammability, modularity, and homogeneous data representation [5•, 8]. The designed and developed homogeneous database and digital platform design is abstract on a technical level and, therefore, widely applicable in architectural terms to other digital platform ecosystems. Thus, we address the call for novel digital platform architecture design [5•] and offer our main contribution through the conceptualization of a dynamic, homogeneous database and digital platform architecture on the one hand, and the data source pipeline to cope with heterogeneous data on the other hand to the literature. Our paper presents the specific case of sleep data as used in the Sleep Revolution, while we would like to argue that the architecture is generic and can be generalized to different healthcare research and might even fit the purpose of research in general.

Our database and digital platform architecture supports heterogeneous data from diverse sources of sleep data and thus has the potential to support multi-disciplinary research needs by effortlessly bringing them the data they need in a preferred format. We see large benefits of such architecture, especially for multi-disciplinary research topics such as sleep. Due to the ownership of data, different end-users can access relevant data without being limited to specific data sources, which is an additional design quality that encourages collaboration between the research fields. Furthermore, the ownership makes deleting participants data entries effortless. This is an essential feature for the Sleep Revolution and other research projects, since the data belongs to its participants, and they can withdraw their participation and their data at any time [17].

The design is particularly suited for projects that involve multiple stakeholder groups and various data sources and,

thus, fulfills the needs for a digital research platform as outlined by Arnardottir et al. [4]. The previously mentioned design qualities are not only sought after in interdisciplinary research projects but are also relevant for other application areas that deal with large amounts of data. In contrast to existing architectures such as the modular layered architecture by Yoo et al. [7•], our design is of a technical nature. This way, we narrow the gap in the literature for more detailed architecture. Moreover, our design is close to practice and easily applicable. This sets it apart from mainly theoretical findings such as those presented by Bygstad et al. [12]. Our approach also differs from other practical solutions such as the research by Bache et al. [11], as it does not operate as part of the communication with the database but instead tackles the issue on an architectural level.

There are limitations to our dynamic, homogeneous database and digital platform design, e.g., it is a less suitable fit for raw signal value entries. In those situations, direct use of raw data requires additional database tables or a file system as a supplement. However, the design fits metadata on each of the raw signal values such as average, min, max, duration, and more. The metadata values can therefore give the end-user a sufficient representation to understand the raw data's amount and diversity.

We did not eliminate the need for preprocessing of the data, and new data sources still require manual work of crystalizing new data source pipelines in order to translate it into the database design. However, the data source pipelines encourage increased automation which minimizes the work needed to add new data points. The data source pipeline's direct communications with our database, that is, the generate-pipeline and insert-pipeline, are independent components that are shared between all data source pipelines. The independent components have a strict evaluation of the data ensuring consistency and homogeneity for all current and future data sources. Thereinafter, the data is presented in the interface, which outlines the front-end of the digital platform. Further research is needed to find suitable visualizations of the data for different end-user groups of the digital platform. The distinct goals for each user group create unique visualization challenges. A high level of customization allows researchers from different research fields to work with data from selected sources to answer their research.

Conclusion

In this paper, we show the complexity of combining various data sources in sleep research and how the researcher's various requirements can be met with a homogeneous database design in a digital platform. First, we contribute with the conceptualization of a simple homogeneous database

and digital platform architecture that uses five tables to represent all data sources with optional additional information that help end-users understand the collected data. Therefore, the complexity of the architecture does not grow with additional data sources. The shared format makes the process of comparing, collecting, and exporting the data effortless for researchers and developers. Furthermore, sharing the data format can connect different research fields by giving researchers a helping hand in using a data source collected outside their field, presenting it in a feasible manner. The design has participant ownership to each data point, which makes the deletion of the participant's data effortless. That is an essential feature in most research projects since the data belongs to the participant who can withdraw its participation, and therefore, their data.

Our action design research findings, derived from Sleep Revolution, provide an illustrative example to show the flexibility and dynamic capacity of the design. However, we argue that this particular digital platform design can be generalized and utilized in other contexts. In addition to that, we contribute with the data source pipeline, which describes all the preprocessing needed for each unique data source from their heterogeneous source into the homogeneous database and digital platform. The data source pipeline with its four key steps, (i) parameter formulation, (ii) forms and entries creation, (iii) process-pipeline creation, and (iv) data insertion, is an obligatory component for the homogenous database to overcome the data sources format constrictions and transform the data into a homogeneous and valid format.

Our design offers a generic design with a high level of customization, and as argued before, it is therefore not limited to sleep research. Instead, it has the potential to fit other fields that require organizing and bringing together large, complex, and diverse datasets in a dynamic manner.

Funding This research is a part of the Sleep Revolution project, with funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 965417.

Data Availability The paper focuses on database and pipeline designs, and no data from this paper can be shared with other researchers.

Declarations

Conflict of Interest Dr. Arnardottir discloses lecture fees from Nox Medical, Philips, ResMed, Jazz Pharmaceuticals, Linde Healthcare, Alcoa—Fjardaral, Visitor (Novo Nordisk), and Wink Sleep. She is also a member of the Philips Sleep Medicine & Innovation Medical Advisory Board. Mr. Sveinbjarnarson discloses fees from Alcoa – Fjardaral and Reykjavikurborg. The other authors declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent This article does not contain any data from studies with human or animal subjects performed. The authors got ethical approval and informed consent before volunteers tested the digital platform.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

Papers of particular interest, published recently, have been highlighted as:

- Of importance
- Of major importance

- 1.●● Arnardottir ES, Islind AS, Óskarsdóttir M, et al. The Sleep Revolution project: the concept and objectives. *J Sleep Res.* 2022;31: e13630. **This paper reflects the need for designing digital platforms in sleep research.**
2. Pevernagie DA, Gnidovec-Strazisar B, Grote L, Heinzer R, McNicholas WT, Penzel T, Randerath W, Schiza S, Verbraecken J, Arnardottir ES. On the rise and fall of the apnea–hypopnea index: a historical review and critical appraisal. *J Sleep Res.* 2020;29: e13066.
3. De Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exerc.* 2019;51:1538–57.
4. Arnardottir ES, Islind AS, Óskarsdóttir M. The future of sleep measurements: a review and perspective. *Sleep Med Clin.* 2021;16:447–64.
- 5.● de Reuver M, Sørensen C, Basole RC. The digital platform: a research agenda. *J Inf Technol.* 2018;33:124–35. **This paper outlines the importance of novel architectural design to cope with increasingly diverse data.**
6. Islind AS. Platformization: co-designing digital platforms in practice. Trollhättan: University West. PhD Thesis, University West. 2018;25:123. Available from: <https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1238297&dswid=-9964>.
- 7.● Yoo Y, Henfridsson O, Lyytinen K. Research commentary—the new organizing logic of digital innovation: an agenda for information systems research. *Inf Syst Res.* 2010;21:724–35. **This paper outlines the need for strategic thinking in designing digital infrastructures.**
8. Constantinides P, Henfridsson O, Parker GG. Introduction—platforms and infrastructures in the digital age. *Inf Syst Res.* 2018;29:381–400.
9. Wareham J, Paul FP, Cano Giner JL. Technology ecosystem governance. *SSRN Electron J.* 2013. <https://doi.org/10.2139/ssrn.2201688>.
10. Nikayin F, de Reuver M. Opening up the smart home: a classification of smart living service platforms. *Int J E-Services Mob Appl.* 2013;5:37–53.
11. Bache R, Miles S, Taweel A. An adaptable architecture for patient cohort identification from diverse data sources. *J Am Med Inform Assoc.* 2013;20:e327–33.
12. Bygstad B. Generative innovation: a comparison of lightweight and heavyweight IT. *J Inf Technol.* 2017;32:180–93.
13. Sein MK, Henfridsson O, Purao S, Rossi M, Lindgren R. Action design research. *MIS Q.* 2011;35:37–56.
14. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational

- research informatics support. *J Biomed Inform.* 2009;42(2):377–81. <https://doi.org/10.1016/j.jbi.2008.08.010>.
15. Carney CE, Buysse DJ, Ancoli-Israel S, Edinger JD, Krystal AD, Lichstein KL, Morin CM. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep.* 2012;35:287–302.
 16. Schmitz L, Sveinbjarnarson BF, Gunnarsson GN, Davíðsson ÓA, Davíðsson ÞB, Arnardóttir ES, Óskarsdóttir M, Islind AS. Towards a digital sleep diary standard. *MCIS 2022 Proceedings.* 2022;10. https://aisel.aisnet.org/amcis2022/sig_health/sig_health/10.
 17. Radley-Gardner O, Beale H, Zimmermann R, editors. *Fundamental Texts on European Private Law.* Oxford, United Kingdom: Hart Publishing. 2016. <https://doi.org/10.5040/9781782258674>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.