JOHANNES KEPLER
UNIVERSITY LINZ

Submitted by
**Lisa-Marie Huber, BA MA BSc**

Submitted at
**Institute for Telecooperation**

Supervisor and Evaluator
**Univ.Prof.in Dr.in Gabriele Kotsis**

October 2022

# A REFERENCE FRAMEWORK FOR EVALUATING VIRTUAL CONFERENCES

Master Thesis

to obtain the academic degree of

Diplom-Ingenieurin

in the Master's Program

Computer Science

**JOHANNES KEPLER UNIVERSITY LINZ**
Altenberger Straße 69
4040 Linz, Österreich
www.jku.at
DVR 0093696

JOHANNES KEPLER
UNIVERSITY LINZ

# Abstract

Virtual conferences not only serve to disseminate new scientific knowledge, but are also intended to promote social networking and sustainability. The aim of this master thesis is to develop a reference framework for the evaluation of virtual conferences with a software engineering-based approach. By means of a meta-study, evaluation criteria and corresponding metrics are defined. A table of sample metrics was also designed to guide the selection of appropriate metrics. Furthermore, an evaluation criteria fingerprint was created, with which the characteristics of a virtual conference can be captured at a glance and conveniently compared with other conferences. Based on these results and a requirements analysis, a corresponding reference framework was designed and implemented. With this flexibly extendable tool, virtual conferences can be evaluated on the basis of metrics. As a special feature, custom metrics can also be defined and added to the dashboard, so that the framework can be used for generic virtual conferences in the field of computer science.

JOHANNES KEPLER
UNIVERSITY LINZ

# Zusammenfassung

Virtuelle Konferenzen dienen nicht nur der Verbreitung neuer wissenschaftlicher Erkenntnisse, sondern sollen auch die soziale Vernetzung und Nachhaltigkeit fördern. Ziel dieser Masterarbeit ist ein Referenzframework für die Evaluierung von virtuellen Konferenzen auf Basis eines Software-Engineering-Ansatzes zu entwickeln. Mittels einer Metastudie werden Evaluierungskriterien und entsprechende Metriken definiert. Als Orientierungshilfe für die Auswahl an passenden Metriken wurde auch eine Tabelle mit Beispielmetriken entworfen. Weiters wurde ein Evaluierungskriterium-Fingerabdruck kreiert, mit welchem die Charakteristiken einer virtuellen Konferenz auf einen Blick erfasst und komfortabel mit anderen Konferenzen verglichen werden. Basierend auf diesen Ergebnissen und einer Anforderungsanalyse wurde ein entsprechendes Referenzframework entworfen und implementiert. Mit diesem flexibel erweiterbaren Tool können virtuelle Konferenzen anhand von Metriken evaluiert werden. Als eine Besonderheit können dazu auch eigene Metriken definiert und zum Dashboard hinzugefügt werden, sodass das Framework für generische virtuelle Konferenzen im Bereich Informatik verwendet werden kann.

JOHANNES KEPLER
UNIVERSITY LINZ

# Contents

# 1 Introduction

Virtual conferencing has become the dominant format for conferences during the COVID-19 pandemic. This necessity to move to virtual conferences, if not to forgo holding the conference altogether, provided an incentive for further exploration of holding formats other than the established physical conferences. Already in the 2020 article "A year without conferences? How the coronavirus pandemic could change research," Giuliana Viglione [Vig20] encouraged researchers to use this extraordinary situation as an impetus to look for better ways to hold conferences. Criticisms highlighted by Viglione [Vig20] were the high carbon footprints, not least due to long-distance international flights, a lack of accessibility for disabled attendees, and the exclusion of potential participants due to high travel costs or time constraints.
In 2021, Williams et al. [Wil+21] published an article on the environmental impact of research communities, pulling together insights from ACM conference sustainability chairs. In light of the global climate crisis, it is of particular interest that the scientific community contributes to carbon dioxide reduction and leads the way in a pioneering role. This paradigm shift also requires traditional ways of holding conferences. However, finding ways to evaluate and compare conferences is even more important so that improvements in carbon footprint reduction become measurable and visible. Currently, the research in this field is lacking, according to Williams et al. [Wil+21].
Fulcher et al. [Ful+20] criticize that conference assessment and evaluation are not sufficiently researched. Virtual conferences have many merits. However, according to Fulcher et al. [Ful+20], some shortcomings only become apparent after the conference has been held through analysis of collected statistics, self-assessments, participant surveys, and informal attendee feedback. Deficiencies were particularly apparent in social networking and technical connection quality [Ful+20].
For this reason, we see the need to focus on evaluating virtual conferences as a research topic and aim to provide a suitable reference framework for these tasks.

## 1.1 Research Questions

Based on the evidence from the literature, we formulated three research questions.

**RQ1** What is the anatomy of a conference, and how can it be visualized adequately with all parts and features?

**RQ2** What metrics can be used for evaluating virtual conferences?

**RQ3** How can we collect and compare the necessary data?

The research questions related to the three-dimensional axis of the cube in Figure 4.1, which comprise the following ranges:

- As a first dimension, the granularity of the core anatomy components of a conference defines the units under test for every evaluation.
- The second axis comprises the number of aspects or metrics to represent all conferences and not be too overloaded.
- Closely coupled to the second dimension, the third one determines the amount of data that can and is allowed to be automatically collected due to technical and legal restrictions. It also questions what data is helpful to collect automatically for evaluating virtual conferences.

The cube and its axis will be explained in more detail in Section 4.

## 1.2 Structure

The structure of the master thesis is as follows: After the introduction, written by the whole team, there is an introduction to evaluation techniques in Section 2. This pre-analysis is written by Alexander. The following section contains an in-depth analysis of existing literature in Section 3 and a survey of the current state of conferences, contributed by Alexander until Section 3.2. The rest of Section 3 is written by Lisa-Marie and includes all characteristics and problem areas of physical and virtual conferences. This lays also the ground work for the aforementioned cube in Figure 4.1. Afterward, the main part of the master thesis begins in Section 4, in which the research questions are dealt with based on three corresponding dimensions, and the basis for an evaluation framework is created. In this in-depth analysis and response to RQ1, conferences and their core components and core processes are first identified and illustrated in a use case diagram, see Section 4.2, written by Alexander. This analysis is followed by the response to RQ2 in Section 4.3. Evaluation criteria and metrics matching them are defined and collected in a tabular overview; see Table 4.1. Section 4.4 will then

discuss techniques for collecting data and highlight challenges and risks, particularly concerning the GDPR or collecting data for so-called green metrics. Both Section 4.3 and 4.4 are contributed by Lisa-Marie. In this section, an evaluation criteria fingerprint is also presented. It can be used to compare virtual conferences and gain insights into their individual characteristics at a glance. After an in-depth requirements analysis and a comparison of already existing evaluation frameworks, Section 5 features the design and development process of the reference framework and details its technical implementation, done by both. RQ3 is answered in Section 4.4 and 5 together. Finally, we conclude the master thesis in Section 6 with results, a discussion and a future work part, written by Lisa-Marie.

## 1.3 Motivation for Developing a Reference Framework

First, an essential question arises: Do we need such a framework, and is not some application already out there that satisfies our needs? We will show in Section 5.1, most of the reviewed tools, on the one hand, only focus on the performance aspect of the used software [Cor20; Ver22; Bar+21]. On the other hand, Read.ai [AI22] indeed opens up the criteria but collects more data than necessary. However, virtual conferences are often much more than a central meeting that can be monitored, and comparable metrics can be calculated from the observations. In addition to various styles and program elements during a conference, the event itself is perceived quite differently depending on the participant's role. This role distinction is not accounted for in the previous solutions. This role distinction means that a presenter or chair focuses on different aspects, thus, shifting the often subjective opinion in opposite directions. Therefore, Section 4 tries to navigate towards an evaluation framework that uses a reasonable amount of metrics that allow for comparing different flavors of virtual conferences and aids in evaluating the conference from different viewpoints.

## 1.4 General Terminology

In this thesis, we will use the term conference as a broad and general term to describe similar events, such as colloquiums, panels, round-tables, seminars, symposiums, and workshops. Although the focus of this master thesis is on conferences in the field of Computer Science and from the academic environment, the results can also be used for conferences in other domains.

Traditionally, a conference is held at a physical location, where research contributions in the form of papers are presented to a professional audience, followed by a discussion on the presented matter. However, as a recent development, conferences are increasingly hosted

online as a hybrid event with both physical and virtual aspects. E.g., a conference that is held in presence but also streamed online or as a solely virtual conference, in which all attendants connect remotely to the hosted conference.

Even though the dissemination of new scientific knowledge is one of the critical aspects of conferences, other vital aspects include establishing and developing social networks among scientists, face-to-face interaction, and the chance to discuss scientific matters with other professionals [SFS19]. Therefore, we define the term conference in the following way:

> *A conference is an event held for disseminating new scientific knowledge among researchers, as well as interaction and social interconnection with a professional audience.*

With this definition, we want to emphasize the importance of interactive and social aspects as a vital part of a conference. To specifically distinguish a conference that is held purely virtually from a conference in the general sense, we need to adapt our definition to include this aspect. Organizational modes such as physical, hybrid, or virtual are omitted so conferences can remain a general term to encompass different kinds of specific conferences. Therefore, we define the term virtual conference in the following way:

> *A virtual conference is a type of conference hosted on the internet and remotely joined by the attendants.*

Additionally, we would like to point out that virtual conferences are sometimes referred to as online conferences in the literature. To distinguish a conference that is only held in a specific location from a conference in the general sense, we need to adapt our definition to include this aspect. Therefore, we define the term physical conference in the following way:

> *A physical conference is a type of conference held in one or multiple physical locations with live presentations of scientific research and in-person attendance.*

To distinguish a conference that is partly held in a specific locale and partly hosted on the internet from a conference in the general sense, we need to adapt our definition to include this aspect. Therefore, we define the term hybrid conference in the following way:

> *A hybrid conference is a type of conference that is held in one or multiple physical locations with live presentations of scientific research and in-person attendance, combined with one or more online-based features, e.g., streaming, interaction with an online audience, or virtual-only sessions.*

A conference in the field of Computer Science may involve multiple so-called tracks that consist of multiple sessions. Tracks enable the parallel presentation of submitted papers and offer the attendants a diverse program from which they can choose the topics that particularly interest them. Tracks are not necessarily related to one conference, as multiple conferences can

be held conjointly. In this case, some of the tracks could be reserved for specific conferences only.
A session is usually dedicated to a particular topic and consists of one or more presentations that directly relate to that topic. Sessions are hosted by a session chair which functions as an expert on the topic and moderator of the presentations and discussions, typically by contributing at least one question for the presenter.
In contrast to classical paper presentations in sessions, a keynote or invited talk is usually the first of so-called plenary sessions. A keynote is a highly anticipated presentation, typically on a topic that contains relevant insights for the entire professional audience at the conference.

In plenaries, most participants gather in one hall/meeting to listen to one or more of the key thinkers in a specific area and be aware of their work [Pop18]. Speakers can be respected persons in a field who do not have to be among the top writers or researchers. We define this term as follows:

> *A keynote is a plenary presentation in conferences that give insight into the work of respected speakers in a specific research field, even outside of the esteemed top researchers.*

As another type of plenary session, a panel is divided into multiple presenters. Often, there are three to five shorter presentations in the same time slots, which are grouped thematically. Depending on the topic, a moderated discussion is started after each panelist's introduction and presentation. Comments on the other presentations and questions from the audience can be topics. Thus, we define this term in the following way:

> *A panel is a plenary session often consisting of multiple speakers in a related thematic and an optional ending with a subsequent discussion.*

Popovic [Pop18] associates a poster or short presentation with a reception. Taking place over mostly 60-90 minutes, attendants can walk around a hall and view the different exhibited posters. At some conferences, a bell rings, encouraging the participants to move around and talk with all authors if possible. Usually, there are design guidelines for creating posters. Effective ways make the best use of visuals and limit blocks of text. Also, interactive components are pretty beneficial, as these sessions are an excellent possibility to talk to many conference attendees and deliver an adequate and clear take-home message.

A call for contributions is a procedure to invite researchers to contribute scientific talks and papers on a specific topic to a conference. It usually includes detailed requirements of how the contribution should be structured and submitted. However, the conference's aims and format should also be considered to determine whether a paper is suitable for a conference. An accepted contribution usually requires one of the authors to give a presentation during the conference.

The best paper/contribution/presentation award is an honorary ceremony, mostly held at the closing of a conference. Its purpose is to recognize and reward quality scientific contributions. Depending on the conference, there are restrictions regarding the eligibility, like submissions co-authored by committee members. Nevertheless, in general, most accepted submissions are eligible for an award. After the responsible committee selects the winner, the authors receive recognition, e.g., a signed certificate or a mention on the conference website.

At this point, we would like to point out that there are many more types of sessions, and this is just a representative sample of Computer Science conferences. Other possible sessions could be, e.g., tutorials, hackathons, and Ph.D. schools. Our reference framework, therefore, needs to be extensible and flexible in design to accommodate this variety in conferences.

## 1.5 Conference Attendants

In this section, different types of conference attendants are described, and their motivations and roles are detailed. This introduction is vital for defining user roles that are essential components in the use cases needed for the first axis of the cube. Section 4.2 discusses these defined use cases in detail.
We applied the same line of thinking to formulate a definition for a conference. We want to broaden the general definition of a researcher to include social aspects and emphasize that research is not an isolated activity. Another central point to stress is that a researcher should also be participating in disseminating knowledge by spreading it to academic peers to advance scientific knowledge in a field. The knowledge should be available to the public and not solely distributed among peers. Therefore, we define this term in the following way:

> *A researcher is a person that is striving to advance knowledge on a certain subject by studying it with science-based techniques, interacting with academic peers, and contributing to the dissemination of knowledge to the public.*

Speakers are commonly selected by a call for contributions procedure or by invitation. A speaker will present the submitted paper the committee has accepted and selected in a session. Additionally, a speaker will try to answer any questions that arise during the presentation.
A chair is in charge of the moderation of a session. Usually, a chair is a renowned scholar and expert in the field to which the topic of a session relates. Figure 1.1 shows a simplified process of becoming a conference speaker in a flow chart to help fit defined terms into the bigger picture of a (physical) conference.

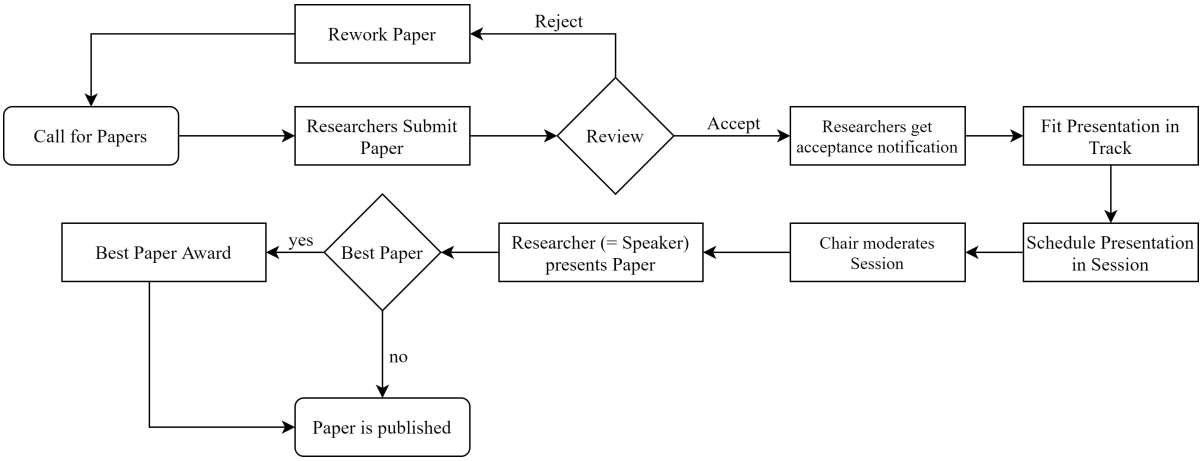



**Figure 1.1:** Flow Chart of a typical conference from a researcher/paper view

# 2 Evaluation Techniques

In this section, more in-depth information about evaluation and related terms is given.
Even before an analysis, some basic terms have to be defined. First of all, what even can be used for comparison? The specific research field in which comparison and assessment are categorized is *evaluation*. It is an important and frequent catalytic activity, which adds significant value to projects [DF08]. Successful evaluation provides insightful data that can be utilized to make informed decisions. Before approaching different kinds of evaluation and which ones to employ for this thesis, it is essential to know why time and effort should be expended for evaluation. Dobrovolny and Fuentes [DF08] see three main reasons to invest in it. First, a thorough evaluation provides a clear justification for performance improvement interventions. Strategies and practices also demonstrate connections with business goals.
Furthermore, organizational decision-making and planning can be improved. This fact is further underlined by the inclusion of evaluation into one of the ten standards of Human Performance Technology (HPT). Systematic evaluation of the efficiency and effectiveness of all interventions is needed to compare the costs incurred and the benefits gained. It is also an essential strategy for advancing and improving a profession.
Their third reason is to provide a strong foundation for continuous improvement responsibilities. HPT practitioners are frequently challenged to provide quality solutions with few resources. Evaluation helps in making decisions smarter. These decisions range from expanding effort to ensuring continuous improvement efforts' effectiveness. Evaluations based on a written and approved strategy integrated into an organization's mission, vision, and operating procedures can become the core of a routinized, efficient, and continuous improvement philosophy.

In literature, defining evaluation is a feared activity. Dobrovolny and Fuentes [DF08] think this might be the case because practitioners fear results that point to a failure or will be negatively scrutinized, especially from inside their organization. In contrast, HPT practitioners have a professional obligation to provide unbiased, sound evaluation data to help leaders make intelligent decisions.

## 2.1 Qualitative Versus Quantitative Evaluation

Evaluation is split into two major categories:

- Qualitative methodologies
- Quantitative methodologies

Although there is a distinction between both categories, they share a standard set of characteristics. Dobrovolny and Fuentes [DF08] state that both are based on a conceptual framework, explanation, rationale, or theory. They also contain straightforward evaluation questions and an evaluation plan congruent with and able to answer those questions. Another vital aspect is employing a systematic approach to the evaluation process and following established guidelines, which have been tested and refined over many years.
A key factor is the data- or information-gathering process for measuring. In that context, it is crucial to protect the collected data, as participants are usually granted anonymity. In exceptional cases, data leaks might lead to complete deanonymization, especially when much personal data is required to collect.
However, it has to be noted that both quantitative and qualitative evaluation can involve judging and decision-making. Therefore the perception of participants plays a central role [DF08]. Question answering and results can be perceived as subjective, controversial, and emotionally or politically charged.
Apart from all these similarities, there are also some main differences in both methodologies. The central separation of both methods is shown in Table 2.2.

However, when to use which methodology? The following paragraph discusses the advantages and limitations of both methods. First, both methods allow us to challenge assumptions, answer an important question about the topic of interest, and question the status quo. Following that, the exploration of issues in a systematic way is supported. However, evaluation may produce politically charged data. Sometimes, an evaluation routine may create more questions than answers. Furthermore, the most crucial thing that limits success: an evaluation is of no help if the wrong questions are asked.
On the one hand, we have quantitative methods that are efficient to execute concerning time and resources. Because of the lower amount of needed resources, human interaction is limited. Due to the generalist approach, results may also apply to a large group of persons. Also, anonymity is easily possible, especially with an automatic collection. However, these evaluation methods answer the questions but neglect substantive or peripheral ones. Thoughtful planning is required for success in this area. Moreover, it has to be noted that a more significant number of evaluation subjects is needed for a precise result. Therefore at least 20 to 50 participants are required.
On the other hand, qualitative methods work with rich data. Therefore, the "why" and "how"

| Quantitative Methods | Qualitative Methods |
|---|---|
| Seek to validate whether a particular assumption or hypothesis is true for a given context. | Focus on context. The results emerge from what is naturally existing in that context. |
| Assume an objective reality that is relatively constant (positivist perspective). | Assume that individuals create their reality independently and socially (constructivist perspective). |
| Separate and detach the observer from the observed. | Involve the observer and the observed, often creating a participant-observer role. |
| Explore population characteristics or sampling frames that represent population characteristics. | Study individual cases or groups that may not represent a larger whole. |
| Refer to the people who participate in the research as subjects. | Refer to the people who participate in the research as participants. |
| Randomly select samples that are as large as possible. | Select participants based on specific characteristics, i.e., a "purposive sample" |
| Describe behaviors with numbers. | Describe actions using words or music, for example. |
| Examine behavior and other observable variables. | Examine the meanings that individuals create and other tacit knowledge or behavior. |
| Explore human behavior in natural or experiment-like settings. | Explore human behavior in its usual context. |
| Analyze social reality according to predefined variables. | Situate observations within a context. |
| Use preconceived concepts and theories to determine what data will be collected. | Discover concepts and theories after data have been collected. |
| Use statistical methods and inference to analyze data (e.g., chi-square, ANOVA, regression techniques, and multivariate analysis). | Use induction to analyze available data (e.g., code interview transcripts, identifying themes and patterns). |
| Generalize findings from a sample to a defined population. | Do not seek to generalize findings unless similar cases exist; instead, provide a detailed description. |
| Prepare impersonal, objective reports of research findings; the final report typically contains charts, graphs, and tables that summarize the data. | Prepare a discourse-intensive final report that describes the themes and patterns and supports those themes and patterns with exemplary quotations or stories. |

**Table 2.2:** Comparison of characteristics of quantitative and qualitative evaluation methods, slightly summarized from the table provided by Dobrovolny and Fuentes [DF08]

of evaluation results are now relevant. Also, they develop a deeper understanding and rapport with the participants. Unfortunately, they come with three significant downsides. They begin with the generalizability of results, which is very difficult. Also, it depends on many factors, namely the questions, the diversity of participants, and the consistency of results. Especially with manual interviews, the process of executing a qualitative evaluation is very time-consuming and dependent. And most importantly, in contrast to quantitative methods, this methodology is resource intensive.
With these facts in mind, the rest of this thesis leans towards quantitative evaluation methods. In evaluating virtual conferences, the focus is shifted to comparing metrics that can be collected as automatically as possible. Energy consumption or performance measurement, like in the analysis by Bieringa et al. [Bie+21], is often done in the context of clients or households and is generalizable. However, conferencing software can also be used for qualitative evaluation, which the paper by Archibald et al. [Arc+19] shows. This evaluation was part of a broader study exploring stakeholders' perspectives on frailty and frailty screening and was conducted with the help of Zoom[1] and about the key advantages and disadvantages of using it.

## 2.2 On Rating Scale Notations and Metrics

The field of evaluation is a broad spectrum full of different facets and questions to ask. One of the more common questions regarding an event or a product is most likely: "How do/did you like it?". When asked orally, the response is often not precise or comparable. However, in surveys, such questions are asked with answer possibilities on a mostly 5- or 10-point scale. Such scales are called *Rating scales*. Originating from the field of psychometrics [Jos+15], such a scale was devised to measure attitude in a scientific and validated manner. An attitude is defined as a preferential way of reaction or behavior in a specific situation in a relatively enduring organization of beliefs and ideas tightened through social interaction. Since their introduction by Thurstone and Likert in the late 1920s and 1930s, this method has been among the most important and frequently used instruments in social science data collection [MB16]. Menold and Bogner describe a rating scale as a continuum with the help of which different characteristics and phenomena can be measured in questionnaires. Likert scales, in particular, orientate the dimension of agreement, ranging from (completely) agree/disagree. The answers are perceived as a function of two fundamental characteristics of said scales:

1. Range, delimited by the scales' poles
2. Frequency, determined by the number of response categories

[1] `https://zoom.us` (accessed: September 2022)

Especially the first characteristic is important, as the range provides crucial information, which is essential for a fundamental understanding of the continuum to be measured. Menold and Bogner [MB16] also state that the response quality of a questionnaire is often tightly coupled to the design. Bad design choices might lead to undesirable effects such as acquiescence, also known as the tendency to agree with items regardless of their content. They further see that the task of answering questions should not be too complex or difficult, nor should it unnecessarily tempt the respondents to reduce their cognitive burden.
Based on their research, Menold and Bogner [MB16] give suggestions on proper design rating scales. These seven tips are as follows:

- The number of response categories (points) should be in the range of five to seven. This range can vary, particularly when unique data analysis methods are used.
- The usage of verbalized rating scales should be preferred over numerical scales, as they might benefit people with low to moderate formal education.
- When used, negative numerical labels may produce systematic effects, like more positive responses, and should therefore be avoided.
- No recommendations can be made about the scale orientation, so either ascending or descending order can be used.
- A scale midpoint should be provided at all costs. Respondents with moderate or neutral opinions may systematically distort the data when alternative categories must be chosen.
- Item-specific rating scales should be preferred over Likert-type scales, as they provide more accurate data.
- If graphical elements are used, the representation should reflect the symmetry of the scale and the equidistance between response categories.

These recommendations should be used with caution, as they only reflect the findings of the authors based on the consensus of the reviewed literature. Especially on the topic of scale orientation, there is no data for a strong recommendation present at the time of writing this thesis.

However, as seen in Section 2.1, rating scales are a very subjective evaluation method, as these continuums can only be measured manually with the help of questionnaires and respondents. This fact is a fatal weak point when considering automation and continuous evaluation. For that reason, rating scales are used only as supporting material in the context of this thesis. Metrics are a better fit to be more comparable. Ellingwood [Ell17] defines the term as follows:

> Metrics represent the raw measurements of resource usage or behavior that can be observed and collected throughout your systems. These might be low-level usage summaries provided by the operating system, or they can be higher-level types of data tied to a component's specific functionality or work, like requests served per second or membership in a pool of web servers. Some metrics present a total capacity, while others are represented as a rate that indicates the "busyness" of components.

Collected data is aggregated or transformed to create comparable measurement values to put it in perspective of the information given in Section 2.1. In software, metrics show compliance towards quality attributes, such as complexity, modifiability, or usability [Ven+18]. These can be a lot of simple counting metrics or, as mentioned before, a set of aggregations, like an index for measuring cyclic dependencies for the number of total dependencies and the size of the software system. However, the methods mentioned above are no all-purpose solution, especially when considering more superficial but constantly changing metrics. In particular, for conferences, a good example is the number of participants during a session. On the one hand, with a simple measurement, no information about the session is gained. On the other hand, one aggregation can only retrieve information about a specific aspect, like a minimum, maximum, or average. Therefore, a collection of observations taken at equal time intervals is called a time series [Ill21]. Unlike causal models, time series models are used in statistics to support analyzing past behavior of metrics. This knowledge can be used for prediction or forecast models, as often practiced in machine learning. They can also be used to identify interdependencies between two or more time series, which helps optimize the precision of the process. Thus, time series and metrics complement each other and should be considered for a sound and comparable evaluation result.
Another important fact is the granularity of measuring metrics in a time series. For example, a company's sales time series are more sensible when measured monthly. CPU usage, however, should be measured in more fine-grained intervals to record phenomenons like short spikes.

## 2.3 Key Takeaways

After this detailed introduction to evaluation techniques, we want to highlight the lessons learned with this pre-analysis and how it affects the rest of the thesis. First of all, as stated in Section 2.1, the work of Dobrovolny and Fuentes [DF08] clearly showed us that the foundation of an evaluation framework for virtual conferences should be built upon qualitative evaluation. Archibald et al. [Arc+19] already showed us the feasibility of this procedure when using conferencing software. We, therefore, focus on refining and aggregating qualitative data to produce comparable metrics for virtual conferences. However, we keep in mind that

the generalizability mentioned above is a significant challenge to overcome. Related to that is also the granularity of metrics to collect. But more on that topic in Section 4.
Moreover, we also have to keep the specifics of the term metrics in mind. When we first learned about rating scales, two fundamental characteristics were pointed out:

1. Range
2. Frequency

Both are essential for receiving accurate feedback when in use. However, can we also use that knowledge to produce more comparable metrics? When recalling Ellingwood's [Ell17] definition of metrics, there can be low-level usage summaries of systems or even higher-level types of data tied to specific functionality or work. Most importantly, however, is that metrics are raw measurements observed and collected throughout systems. Comparability can be achieved by trying to enrich measurements from conferences with an acceptable range. This enrichment can be done with a rating from one to five for eco-friendliness as a range. We can then use a frequency of equidistant steps of 0.5 when following the design tips by Menold and Bogner [MB16]. These steps are possible after mapping CO2 emission ranges to them. Further research on that topic is conducted in Sections 4.3 and 4.4.
Lastly, we must remember that past behavior and continuous progress are also essential to consider. These factors show the presence and also the absence of effort put into improving the conference over time. For that reason, metrics should be annotated with a time component to produce a collection of metrics over the years, called a time series [Ill21].

# 3 Analysis of Conferences

To prepare our knowledge foundation and to be able to answer RQ1 correctly, multiple steps have to be performed. First of all, a quick overview of the state-of-the-art should be gained. This goal is achieved by discussing several reports and lessons learned from conferences held in the past two years. The goal is to obtain insights about similarities, positive aspects, and significant problems, which could be circumvented with a physical conference. Based on aspects from these reports and our own experiences, evaluation criteria will be defined in Section 4.3 and applied to both types of conferences. Hybrid conferences are omitted in this analysis, as we decided to identify them by their primary orientation for the scope of this thesis. Therefore, the following section treats a physical conference with live streams or recordings as physical.

## 3.1 Motivational Aspects for Researchers

Before even comparing different implementations of conferences and weighing their advantages and disadvantages, the focus has to be laid on the fundamental reasons why conferences are held in the first place. This section targets the point of view of researchers, as they are the leading group of attendants in this scenario.
Sanders et al. [San+20] define conferences as supportive events for maintaining shared professional identities, addressing developmental goals, building social networks, and maintaining, developing, and disseminating knowledge about the discipline. Also, a generalization of academics and other experienced personnel in different fields of work is introduced with professionals. Therefore, the defined goals are identical for all professionals. Although it is now clear, *what* a researcher might want to achieve at a conference, it is still unknown *why* they attend such events.
In addition to the career self-management framework by King [Kin04] from 2004, Sanders et al. [San+20] proposed a topology of factors from different dimensions influencing researchers' motivation.

As seen in Figure 3.1, there are three flows of influences on developmental goals. In more detail, three main dimensions are dominant in terms of influence: (i) individual career stages, (ii) type and orientation of university, and (iii) national values. These can be a sole influence

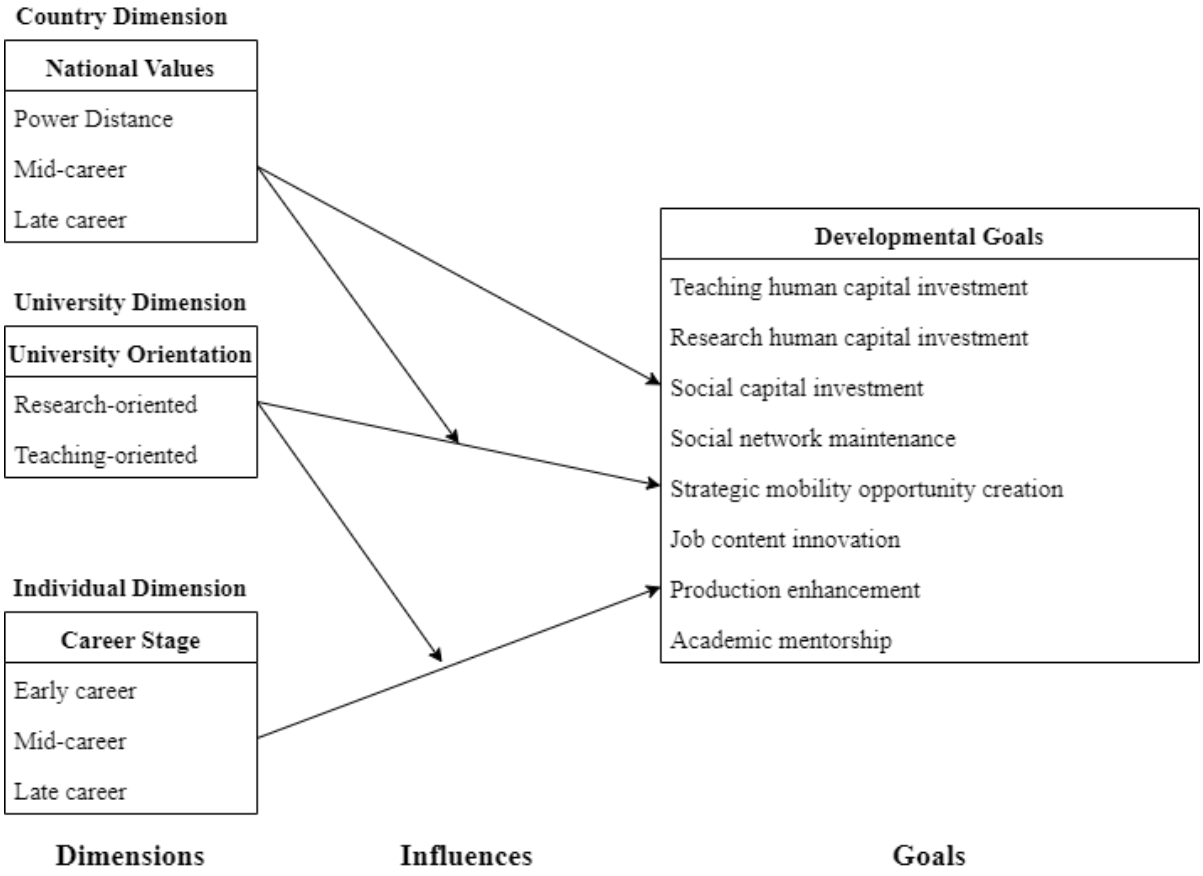



**Figure 3.1:** Theoretical topology of researchers' motivation of attending conferences, restructured from [San+20]

but also (ii) and (iii) can contribute to the flow of (i) and (ii), respectively. But more on that in the respective sections.

### 3.1.1 Career Stage

Beginning with (i), it can be observed that it is very influential at which time of their career the researcher attends a conference. For instance, academics in an earlier career stage are more likely to invest in research and teaching human capital, investing and maintaining social capital, and to achieve productivity enhancement. Human capital enhances one's skills and knowledge in a respective field [San+20]. Research investment is less important than achieving job content innovation when advancing into the mid-career. Thus, it is vital to maintain social networks to consolidate their place as competent contributors to the profession. In the late stages, however, the focus is shifted entirely to maintaining existing networks, creating mobility through strategic job opportunities, and positioning themselves as mentors to other researchers.

### 3.1.2 Orientation of Institution

As for (ii), the academic university's general operational direction substantially influences motivational aspects. Universities and all other higher educational institutes can focus on either research or teaching. In most cases classifications exist for them, i.e. the Carnegie Classification system for the US[1]. Research-oriented institutions, on the one hand, are often granted considerable funds for supporting staff in their ongoing professional development centered around research. Such services include allowing researchers to travel and attend key developmental workshops and conferences to support their research and develop research networks. As stated by Sanders et al. [San+20], the ultimate goal is to help academics develop human and social research capital. On the other hand, teaching-oriented institutions are provided much less financial support to pursue their professional development goals and develop teaching human capital. Examples are the latest innovations in pedagogy and best practices in teaching delivery.
In symbiosis with (i), the university type also influences the developmental goals of different career stages. Early-stage academics are more likely to attend conferences to invest their respective human capital. Compared to research, the teaching-oriented staff is less likely to attend conferences to create strategic job mobility opportunities. In the later stages of their careers, teaching-based academics are more likely to expand their job mobility in their spectrum. However, in the other branch, conferences are often used as a foundation for strategic mobility in research and teaching.

### 3.1.3 National Values

Last but not least, (iii) plays an influential role in the motivational direction of an academic. Important factors are natural culture and power distance. House et al. [Hou+04] define national culture as shared motives, values, beliefs, identities, and interpretations or meanings of significant events resulting from typical experiences of members of collectives transmitted across generations. Sanders et al. [San+20] supplement this statement that culture impacts how members of a nation view learning barriers, solve problems and manage knowledge. More frequently examined is the cultural dimension of power distance. House et al. [Hou+04] define it as the degree to which collective members expect power to be distributed equally. Examples of high power distance cultures include Malaysia, Mexico, and China, where superiors and subordinates consider themselves different from each other, and the first ones are believed to be entitled to privileges [San+20]. In contrast, Scandinavian countries and New Zealand are defined as low power distance cultures, where both parties are seen as equals, and it is believed that no one should be entitled to special privileges.
Regarding academics, the higher the power distance of their nation is, the less likely they

[1] `https://carnegieclassifications.iu.edu/` (accessed: September 2022)

are to attend conferences to invest or maintain social capital. Another interesting factor to consider is in-group collectivism. House et al. [Hou+04] define it as the degree to which individuals express pride, loyalty, and cohesiveness in their organizations or families. Characteristics of high in-group collectivism are countries like India or Turkey, whereas Denmark, Australia, and the Netherlands represent the other side of the spectrum. Finally, the uncertainty avoidance of a nation, which is defined as the extent to which a society, organization, or group relies on social norms, rules, and procedures to alleviate the unpredictability of future events, expresses the degree to which members of a society are comfortable with uncertainty, ambiguity, and risk-taking. Sanders et al. [San+20] utilize these metrics for their propositions. Thus, academics from high in-group collectivistic countries are more likely to attend conferences to mentor other researchers. However, the higher the uncertainty avoidance, the less likely academics will attend conferences to create strategic mobility opportunities and achieve job content innovation.

These national values also heavily influence the other two motivational reasoning dimensions. For instance, high collectivism at the country dimension affects the developmental goals in all three dimensions, so it is more likely in earlier career stages that academics attend conferences for networking, and at later stages, they provide academic mentorship. High power distance, however, shapes the goals of later-career academics. Thus, they are less likely to attend conferences to invest and maintain social networks and vice versa for low power distance countries.

## 3.2 Current State of Conferences

For many, 2020 was a year of abrupt changes, both professional and personal. The outbreak of the Covid-19 pandemic disrupted traditional ways of hosting conferences. It forced the academic community to quickly adapt to virtual conferencing if conferences should not be canceled due to widespread restrictions, especially regarding social distancing [HWK22]. While 2020 needed to switch to virtual conferences, in the second and third years of the pandemic, other formats were possible, even physical conferences in some regions and during specific time frames. However, Hurst et al. [HWK22] found that virtual conferencing entered the mainstream in the academic community and as part of non-academic work environments. Furthermore, they argue that virtual conferencing will become a preferred option for academic conferences since they offer an array of benefits, such as reducing the carbon footprint and lowering barriers to attendance in terms of costs and ease of access. Hurst et al. [HWK22] even conclude that virtual conferencing with a focus on an increased immersive experience for participants may be the ideal solution, even though virtual conferences still prove to be a challenge for the time being. Meyer et al. [Mey+21] pointed out that conference participants

put in additional effort aside from the past paper writing and presentation tasks. Energy for traveling long distances, as well as securing funding for travel, lodging, meals, and registration fees, are secondary factors that they are taking into account for participating in one of the highlights of an academics' year. However, one of the main disadvantages of the past system was tilted towards more privileged scientists, as they often are well-funded, connected, and generally more established in the community.
As opposed to the general idea that might have sparked the community to skip conferences altogether in 2020, this would have come at a cost too high for academics in their early career stages. Referring to Section 3.1.1, participants want to share their work to create job opportunities and establish their first scientific network, as pointed out by Sanders et al. [San+20]. Especially in the aquatic sciences and STEM community, the number of virtual meetings and workshops increased noticeably. Haji-Georgi et al. [HXR21] suggested two main reasons for the rapid increase of virtual conferencing during the COVID-19 pandemic:

1. Virtual conferences were the only possible venue for interactions in an academic setting while social distancing measures were in place
2. The adopters of the virtual conferencing technology, i.e., academics, were homogeneous

Meyer et al. [Mey+21] stated further that the transition to the online mode removed some of the barriers mentioned before. Accessibility for peers unable to participate in past years, which resulted in likely increased participation, was one of the primary positive outcomes of this paradigm shift. Nevertheless, some past hindrances remain, and even new ones arise. The new challenges included access to a reliable network, a meaningful internet connection, and time zone management for conferences with a global audience. Also, the unexpectedly high amount of consumed energy when attending online sessions over a more extended period, and most importantly, finding time for dependent care as many facilities enacted restrictions or limited services due to the pandemic.

Since the topic of this master thesis involves virtual conferences in general, we decided to broaden the scope and not focus only on computer science-related conferences. Therefore, for a further analysis, experiences from six major conferences in the aquatic sciences (GLEON[2], NALMS[3], PPNW[4], EFI[5], KGML[6], and DSOS[7]) [Mey+21], as well as the Neuromatch Conferences [Ach+21], EuroSys 2020 [Bil+20], EDBT/ICBT 2020 [Bon+21], COPERNICUS Alliance

[2] `https://gleon.org/` (accessed: September 2022)
[3] `https://www.nalms.org/` (accessed: September 2022)
[4] `https://ppnw.info/` (accessed: September 2022))
[5] `https://ecoforecast.org/` (accessed: September 2022)
[6] `https://sites.google.com/umn.edu/kgmlworkshop/workshop` (accessed: September 2022)
[7] `https://isdr.org/virtual-summit-incorporating-data-science-and-open-science-techniques-in-aquatic-research/` (accessed: September 2022)

Online Conference 2019 [DZM20], PAM 2020 [Mis+20], and VLDB 2021 [Bon+22] are summarized. Additionally, tendencies are highlighted and personal experiences from DEXA[8] and iiWAS[9] in 2020 and 2021, are fitted into the bigger picture. In the following section, these conferences will be referred as *collection* or *factual base*.
One of the fundamental areas that are decisive for a (virtual) conference's success is the event's modality. Bonifati et al. [Bon+21] set up several questions leading to modal decisions that are helpful for organizers to create an experience tailor-made for the audience. Most of them are similar but introduced in more detail compared to the guidelines provided by Zimmermann et al. [DZM20].

### 3.2.1 Temporal Presentation Aspects

First of all, synchronicity is a crucial factor in setting the general direction of the conference. Whether the event is held synchronously or not can answer some of the following questions in advance. On the one hand, an asynchronous mode with pre-recorded talks and live Q&A sessions might resolve potential time zone issues. Even going further with a communication tool, which allows discussion partners not to be online simultaneously, is another possibility. On the other hand, a synchronous event allows for more live participation and, most importantly, interaction, thus, preserving the spirit of physical conferences. Every conference of the factual base was held in the latter mode, with honorary mentions towards GLEON, PPNW, DSOS [Mey+21], and EuroSys [Bil+20], for additionally supporting the conference with asynchronous chat channels for both discussion and organizational matters.
Time differences are a common problem, directly dependent on this modal aspect, especially when addressing a global audience in live settings. Participants with a stronger desynchronization than others are experiencing the feeling of being detached and handicapped. Across the collection, the consensus was to adapt the time zone of the conference to the primary audience, which is, in most cases, the standard time for the organization team. At DEXA and iiWAS, CEST was used in favor of the organizers at JKU, Linz, thus, resulting in less participation by the Asian audience. At EDBT/ICDT [Bon+21], they opted for shorter conference days, with 5-6 hours instead of 8-9 at physical editions.

Following this temporal aspect is the length of presentations and sessions. EDBT and ICDT shortened their presentation times by roughly 50 % to 10 and 12 instead of 20 and 25 minutes. As a logical consequence, session lengths are not that long anymore, thus resulting in generally shorter conference days. Of course, there are longer talks possible, like plenaries and keynotes. At DEXA and iiWAS, talks had a standard duration of around 15 minutes, whereas other formats had lengths of up to one hour. At EuroSys [Bil+20], they tried to use the knowledge of EDBT/ICDT and asked the authors to prepare short pre-recorded videos of

[8] `http://www.dexa.org/` (accessed: September 2022)
[9] `http://www.iiwas.org/` (accessed: September 2022)

3-5 minutes and real talks of 10-12 minutes. After feedback from the first sessions, the longer videos are presented with live Q&As later. Achakulvisut et al. [Ach+21] can support this result by using a split of 5-minute talk and 10 minutes of discussion per presentation at the Neuromatch Conferences. From the point of view of aquatic research gatherings [Mey+21], data is only available for PPNW with one plenary talk and statistical stray bullets consisting of thirty-minute talks as well as three-minute flash presentations.

At last, the active part of the presentation structure is the shortly mentioned before the topic of recorded versus live talks. Bonifati et al. [Bon+21] think the execution depends on the aim to avoid technical problems as much as possible. When places with optimal streaming capabilities, i.e., stable and fast internet connections, are available and the presenters are not that familiar with video conferencing platforms, pre-recording might be advantageous. However, two additional hindrances arise in connection with live and prepared recordings. First the technical aspect of storage and retrieval, as high-quality videos can become very storage-intensive and break the per-file size limits of some solutions. The second and more important is privacy. In most cases, the authors/presenters are visible in the recordings, so to be GDPR-compliant in the EU[10], some sort of publication consent has to be given. At EDBT/ICDT, live presentations on Zoom were recorded, and the whole collection was uploaded to an sFTP solution. Videos were made publicly available after asking the authors for permission. From the collection, only PPNW, KGML [Mey+21], COPERNICUS [DZM20], Neuromatch [Ach+21], and DEXA/iiWAS held their sessions with a majority of live presentations, mostly over Zoom. All of the others opted for pre-recordings. In general, recordings of at least keynotes and panels were published afterward. Consent was given explicitly by the authors or with the registration with a chance to opt out before publication. For VLDB 2021, Bonnet et al. [Bon+22] decided to have every speaker send in pre-recorded videos of their talks beforehand but encouraged them to hold their presentations live. Despite their encouragement, most speakers opted for streaming the pre-recorded video even if they were present.

### 3.2.2 Program Schedule Decisions

After the first modal decisions, which impacted the general time frame of the conference, the next ones are important for the diversity of the program schedule. As a bridging topic to the former section, managing questions is a fundamental problem to resolve. Bonifati et al. [Bon+21] think that there are two directions of doing so:

1. live questions, whether textual or face-to-face (via voice chat and camera)
2. asynchronous communication channels like messenger services or forums

[10] `https://eur-lex.europa.eu/eli/reg/2016/679/oj` (accessed: September 2022)

At EDBT/ICDT, both possibilities have been used, with a more informal approach when asking questions during the sessions. Here participants had to switch on their cameras and raise their hands to replicate the feeling of a real conference. During keynotes, Zoom Webinar[11] was used, so therefore, the most popular questions were voted and then asked. Additionally, Slack[12] was used with distinctive channels for each session. At the other factual base conferences, all aquatic conferences [Mey+21], EuroSys [Bil+20], and DEXA/iiWAS used Zoom for live Q&As. EuroSys, GLEON, PPNW, DSOS, and PAM also installed an asynchronous message channel with Slack. The latter even used Google Forms to collect questions to send to the authors, which was possible before the respective session. At VLDB 2021 [Bon+22], the event management software Whova was the only entry point for remote attendees to the virtual component of the conference. Whova allows for customized interactions with participants and to continue discussions beyond the duration of the sessions. However, it was decided to attribute a higher priority to interactions between physical attendants and live interactions instead of encouraging asynchronous communication.

The following paragraph summarizes the program decisions from the factual base conferences regarding conference program points apart from presentations, namely short/poster papers, demos, tutorials, and keynotes. First of all, short papers as they are shorter than regular papers and are usually grouped in a poster session more interactively. Bonifati et al. [Bon+21] tried to play videos of all 26 posters back-to-back with a live Q&A at the end. PAM [Mis+20] reorganized the poster session as 5-minute talks with an additional 1-minute Q&A. In the collection, only three of the aquatic conferences [Mey+21] held a poster session, with GLEON trying out an asynchronous way via VoiceThread[13]. Neuromatch [Ach+21] found that interactive talks of 5 minutes with a 10-minute discussion afterward gave the best interaction rate out of their conferences. Similarly to poster sessions, demos are in the same spot. At EBDT/ICDT [Bon+21], 15-minute timeslots were reserved, in a reversed distribution, resulting in longer videos. For demonstration and poster sessions, Bonifati et al. [Bon+21] suggest careful planning to achieve a similar experience compared to physical conferences. Suggestions are breakout rooms in Zoom or the future, more sophisticated solutions based on virtual reality or avatar-based video. As there were only demo sessions with live questions afterward at NALMS [Mey+21], no further discussion is added here.
With tutorials and keynotes being longer sessions and at least the latter more likely to be present in conferences, they generally attract higher attendance. A crucial tip by Bonifati et al. [Bon+21] is to plan more discussion time as everyone feels close to the presenter, and thus, people seem more inclined to ask questions. In the collection, most of the conferences organized at least keynotes or plenary sessions in a pre-recorded or live manner, with EuroSys [Bil+20] being the only one without such a program point. However, only experiences

[11] `https://support.zoom.us/hc/en-us/articles/200917029-Getting-started-with-webinar` (accessed: September 2022)
[12] `https://slack.com` (accessed: September 2022)
[13] `https://voicethread.com/` (accessed: September 2022)

from DEXA/iiWAS told something about the duration of such sessions, with one to one and a half hours per keynote, including introductory words and discussion. Held during the second year of the pandemic, the organizers of the VLDB 2021 [Bon+22] decided to keep the traditional format of VLDB with plenary sessions and up to seven parallel sessions. They did not opt for repeating sessions not to dilute the number of participants present at the sessions.

And last but not least, a difficult, but especially in times of social distancing important topic: the occurrence of social events during the conference. Referring to Section 3.1, networking is one of the crucial activities in an academics' daily work life. The interdisciplinary exchange generates many opportunities and solutions during coffee breaks or hallway chat, which make conferences, in general, more than a success, even more than the presentations themselves [Bon+21; Ach+21; San+20; Bil+20; Mey+21]. Following the factual base, some possibilities can help engage more with the community. Bonifati et al. [Bon+21] tried to cover the conventional social events with joint activities during the coffee breaks. "Bring your own beer" receptions as well as a best-effort alternative for the conference dinner by sharing recipes for home cooking with simple ingredients, which were possible even in the early stages of the pandemic[14], were tried out. They also suggest planning more time than 15 minutes for coffee breaks to make conference days less exhausting. In the aquatic conferences [Mey+21], breakout groups on particular topics, post-session social hours, and scavenger hunts, trivia nights were arranged. The Neuromatch conferences [Ach+21] tried to cover this with their poster session and coffee breaks alone. At EuroSys20 [Bil+20], the organizers created virtual meeting rooms and hosted discussions during the breaks, all in the form of a hallway track on Discord[15]. At VLDB 2021 [Bon+22], where live interactions between physical attendants were prioritized, the setup with Whova to boost the interactions between all participants, regardless of physical or virtual attendance, was found lacking. For the future, the organizers of VLDB 2021 recommend making an extra effort to include virtual participants and utilize a better-suited communication platform like Slack.

### 3.2.3 Organizational Issues

In contrast to the rest, the last points are less impactful for the audience than the organizers themselves. First of all is the chairing of sessions, especially in live settings. Bonifati et al. [Bon+21] initially started with no session chair but only a Zoom host. While this seemed like a successful initiative, this single host was decked out with tasks like, for instance, playing videos and monitoring the chat, all in parallel to chairing the session. An effective strategy,

[14] `https://edition.cnn.com/2020/03/18/business/supermarkets-rationing-coronavirus/index.html` (accessed: September 2022)
[15] `https://discord.com/` (accessed: September 2022)

also used at DEXA[16]/iiWAS[17] is to have a separate session chair, who is part of the community in the respective track branch, to chair the session and have a second host run the session from a technical point of view. It is recommended from experience that this second person acts as some sort of "technical organizer," so the following characteristics are essential:

- Ability to solve technical problems or at least be able to give enough support
- Communication with the session chair and organizers regarding no-shows, ...
- Management of participants, including assigning presentation rights, muting disturbing attendees, ...

From our factual base, all other conferences support this recommendation [Ach+21; Bil+20] to include a technical host per session. In the aquatic conferences, there is no explicit information about the sessions. Nevertheless, Meyer et al. [Mey+21] suggest having at least one technical officer present to set up the meetings and support emergencies correctly. At PAM 2020 [Mis+20], a technical chair was assigned to prepare to stream and manage to host the online sessions. Bonnet et al. [Bon+22] also appointed a digital platform chair to manage both the content of the platform for the virtual component of the conference as well as coordination with virtual attendees.

Risk management is vital for a successful event when speaking of technical difficulties. Following the motto "The show must go on "[18], precautions have to be met to ensure a smooth experience for the conference audience. Bonifati et al. [Bon+21] stated that testing before the event is key. Especially with virtual meetings, corner cases have to be inspected. Disconnects can happen to anyone at any given time. In the case of EDBT/ICBT, when losing the host, the meeting can continue, but the host's video freezes. The main takeaway has backup hosts with at least a crash course on handling the software tools. Also, a guide with general information for hosts, chairs, and technical staff is very beneficial. In the collection, very little detail regarding handling technical problems is spread. Neuromatch [Ach+21] conferences had a live chat backbone via Slack to address all kinds of issues, from last-minute cancellations to no-shows and technical problems. At EuroSys [Bil+20], student volunteers assisted the conference as technical co-hosts and supporting staff in handling problems. They also performed significant training for all Zoom-related staff, even for the session chairs, as most are not familiar with hosting a Zoom session either. This experience overlaps with DEXA/iiWAS, as many chairs put their trust and responsibilities into the technical co-hosts. Zimmermann et al. [DZM20] also underlined the preparation aspect.

[16] `http://www.dexa.org/` (accessed: September 2022)
[17] `http://www.iiwas.org/` (accessed: September 2022)
[18] `https://dictionary.cambridge.org/dictionary/english/show-must-go-on` (accessed: September 2022)

## 3.3 Physical Conferences

To distinguish the features that a virtual conference excels at, we must also investigate physical conferences, as they have been the predominant mode of conferences for years. Neugebauer et al. [Neu+20] state that physical conferences are significant and irreplaceable for knowledge exchange, networking opportunities, scientific discussion, and the creation of new projects, ideas, and solutions in nearly every discipline of scientific research. Attendance at physical conferences is generally perceived as contributing to the advancement of a researcher's career, e.g., by creating learning and networking opportunities and providing a platform for scientific exchange [Roo+20; RSP21]. Experiencing and interacting with renowned researchers or being part of an educated and like-minded crowd of peers is an important part of a researcher's conference attendance [SFS19]. Sá et al. [SFS19] argue that a researcher's visibility cannot be increased in the same way in virtual conferences as with physical conferences. A major concern with virtual conferences is that they cannot adequately emulate these social dimensions of a physical conference. Physical conferences provide more opportunities for bidirectional information exchange. Compared to the more unidirectional knowledge dissemination approach of virtual conferences [SFS19]. Especially informal interactions or the impression of belonging to a community are not easily facilitated in a virtual setting [SFS19]. Sá et al. [SFS19] stress that many of the interactions at a conference occur during breaks, social events, or even in the corridors. Forming lasting professional relationships is also more likely to happen at physical conferences [SFS19]. However, we need to point out that introverted people or people who have not mastered the primary conference language might not be able to reap the benefits of physical conferences in the same way other participants can [SFS19]. These deficiencies may decrease job satisfaction and contribute to burnout and stress [SFS19; RSP21]. Apart from these scientific or career-related social dimensions, cultural exchange and the benefits of broadening one's horizon by learning about new cultures and places is hampered [Roo+20; Woo+21]. Furthermore, the decrease in social networking capabilities disproportionately affects researchers in their early career stages [Woo+21]. According to Woodruff et al., [Woo+21], the main reasons for this decrease are fewer spontaneous discussions, e.g., in the hallway after a talk, a decrease in interactivity and engagement with peers, as well as a general lack of social component of conference participation. At physical conferences, non-verbal communication is possible, as well as the perception of reactions from the researchers in the auditorium [SFS19]. Additionally, confirming facts with fellow participants is more easily possible at physical conferences than in virtual settings [SFS19]. Opportunities to hone public speaking and presentation skills are also lost since presenters at virtual conferences are more likely to read off their notes or send a video of the recorded talk to be presented in their stead [Woo+21; RSP21]. Lamentably, the ad-hoc collaborations, social networking, and spontaneous gatherings cannot be fully replicated by virtual conferences, and therefore a uniquely valuable factor of physical conferences [Neu+20; Woo+21]. Current

criticism of physical conferences mostly revolves around the high costs of attendance, including the traveling expenses and the high CO2 footprint associated with them [Neu+20; Roo+20; Sar+21]. Despite the increasing interest in the sustainability of conferences, Neugebauer et al. [Neu+20] remark that sustainable conference management is not a common practice. CO2 contribution factors, such as catering, conference rooms, and hotel accommodation, are often neglected [Neu+20]. However, a key contributor to the large CO2 footprint of conferences is the traveling aspect [Neu+20]. Another important issue is the lack of inclusivity in physical conferences regarding gender representation, primary care duties, or teaching duties. Early career researchers lack the funding and recognition that seasoned researchers possess, or generally researchers from countries or universities with low available funds [Roo+20; Ach+21; LMO20; RSP21; Sar+21]. Parncutt et al. [Par+21] argue that hybrid conferences, i.e., mostly physical conferences with added virtual presentations, discriminate against remote attendees and encourage flying to the venue. Parncutt et al. [Par+21] suggest hosting multi-hub conferences that mix physical and virtual interactions and are distributed across the globe to enable participants to travel to a hub near them and enjoy the benefits of both conference types to remedy this issue. Foramitti et al. [For+21] state that virtual conferences provide a more inclusive and safe space to participate. An overlooked aspect is that the properties of physical learning spaces, e.g., acoustics, lighting, noise, color, and seating arrangements, and the immersion into the conference experience by physically attending affect learning [SFS19]. With virtual conferences, researchers may lack the digital literacy to handle the tools for participating and presenting at a conference well [SFS19]. The technical entrance barrier at physical conferences is therefore much lower. Even though it is highly improbable that virtual conferences will fully replace physical conferences, this trend towards more virtuality could prove beneficial for physical conferences. As more virtual conferences are held, and more lessons are learned, it becomes more apparent which physical conferences are lacking and how they could be improved by adding some elements of virtual conferences.

Summarizing this section, we present the pros and cons of physical conferences that could be identified, sorted in descending order of importance:

+ Social networking and opportunities for career advancement
+ Unique immersive experience
+ Informal discussions with peers
+ More bidirectional process of knowledge exchange
+ Low technical barrier
- The high environmental impact/large CO2 footprint
- Lack of inclusivity of members of underprivileged groups

- High costs of attendance and traveling
- Power dynamics/less safe space for participants

## 3.4 Virtual Conferences

Comparing the modalities, holding virtual conferences has several advantages for physical iterations. However, the goal of this section is to try to highlight these but point out their lacking abilities as well. The foundation for this evaluation is again the factual base, focusing on survey results and lessons learned. Beginning with temporal aspects, according to Bilas et al. [Bil+20], the most important factor when attending a virtual conference is time with a clear 70 % of the votes, even before financial and environmental reasons. In contrast, Bonifati et al. [Bon+21] were completely satisfied with a completely synchronous approach, requiring the participants to be present at the conference to experience the talks.
In the survey results by Bilas et al. [Bil+20], the audience's consensus is presentation durations of 11 to 20 minutes. According to a survey by Bonifati et al. [Bon+21], this change of presentation time was received very well. However, experiences at EDBT/ICBT with pre-recorded talks were met with mixed feelings. On the one hand, well-planned sessions happened, with fewer slips of the tongue and less likelihood of running overtime. On the other hand, the dynamics of the talks suffered and tended to be more monotonous. At EuroSys20 [Bil+20], however, most survey-takers liked the pre-recorded videos streamed during sessions, with live Q&A afterward. Meyer et al. [Mey+21] also pointed out that recorded presentations can help bridge time zones and the busy daily schedules of researchers. This effect could be seen at DEXA/iiWAS, where Asian researchers, for instance, were more likely present at the early sessions but missing in the later ones. Capitalizing on that fact, they further recommend establishing communication platforms to help facilitate informal conversations as well as the possibility for further in-depth questions.
On the notation of different platforms, Achkulvisut et al. [Ach+21] point out the importance of the used software tools. Compared to the physical counterpart, the user experience is mainly affected by the chosen system. Thus, the perfect solution would be a platform that offers both interactive user interface elements as well as stable, high-resolution streaming.
In the Neuromatch conferences, Achkulvisut et al. [Ach+21] further state that the aforementioned interactive talks with discussion earned the most interaction out of all of their experiments with the poster sessions but still failed to capture the real poster experience. They see potential in technical solutions regarding proximity-graded sharing of virtual spaces, with tools such as GatherTown[19] or Mozilla Hubs[20], but even more innovation is needed. Bonifati et al. [Bon+21] see potential in these tools. However, one of their projected weaknesses is the

[19] `https://www.gather.town/` (accessed: September 2022)
[20] `https://hubs.mozilla.com/` (accessed: September 2022)

time-efficient overview of the presented material when walking through a room physically, which cannot be reproduced by online tools yet. However, they still see the proper use of such tools, like GatherTown, especially for social events.
Also important are economic points of view. Neuromatch conferences [Ach+21] showed that virtual conferences could be very inexpensive. Their estimated costs for Zoom Webinar at Neuromatch 3 totaled around 4000$, which was collected via the 25$ participation fee. Thus, the conference was put on a sustainable path. Furthermore, location costs are almost non-existent, and reduced costs allow a more diverse attendance. Early-career scientists, such as graduate students and postdocs, were counted as the majority of the attendees. This demographic characteristic can also be seen at EuroSys20 [Bil+20] and is recommended by Meyer et al. [Mey+21]. Haji-Georgi et al. [HXR21] view virtual conferences as technologically challenging, especially in terms of creating an immersive experience for participants. However, virtual conferences significantly reduce the costs of participants and organizers, facilitating the spreading of knowledge among researchers without significant investments. Regarding sustainability, Bonifati et al. [Bon+21] observed that more and more rethinking is happening regarding adopting CO2 plans for conferences. Also, the research community is willing to go online to reduce the CO2 footprint. This development may follow the surprising fact that the conferences exceeded their expectations for virtual conferences.
Porwol et al. [Por+22] identify three main communication features that conferences must have in order to provide the best experience for participants:

- Immersion: Being part of the conference experience
- Presence: Promoting a shared sense of presence, i.e., simultaneously being together at the conference
- Sense of Community: Facilitate networking between conference participants

In terms of immersion, asynchronous virtual conferences need to provide recorded sessions with an additional platform such as Slack or Discord to communicate in organized chats. Synchronous virtual conferences may include attendance at live virtual sessions via different tools. Physical conferences are the gold standard for presence but are limited by time and space. On the other hand, asynchronous conferences allow to relive recorded conference experiences and at least supply a solid vocal presence. Synchronous virtual conferences offer only as much presence as the video conferencing tool, and how well it is moderated and is bound to time. A sense of community is possible for physical and virtual conferences. However, virtual conferences may have better interaction possibilities. Porwol et al. [Por+22] believe that the persistence of recordings enables people to relive the sense of community from when the conference occurred. Additionally, they argue that synchronous virtual conferences help alleviate power dynamics problems by allowing early-career researchers to type their questions instead of needing to unmute and speak up. The following important pros and cons of

virtual conferences could be gained to summarize his section as well as the state-of-the-art evaluation, both in descending order of importance:

+ Easy global participation and low financial requirements supporting early career researchers
+ Lower CO2 footprint
+ Lower costs for organizers (no conference location, catering, ...)
+ With proper training, the conference is organizable by a small team
+ Recordings enable participation of busy attendants and bridge time zones
- Less social interaction with no innovative and satisfactory solution to properly network during the event
- User experience is mainly dependent on the choice of software
- Monotony in recorded presentations hampers the concentration of participants
- Synchronous modality risks lower participation concerning time zones.

After this extensive comparison of both major types of conferences, we have to give an outlook on hybrid events at this analysis point. This teaser is especially important, as current trends point in the direction of conferences with at least one concurrent physical and virtual component. Bonnet et al. [Bon+22] and Porwol et al. [Por+22] argue that this trend towards hybrid conferences will likely continue until the majority of events are held in this format.
What does this mean for the pros and cons of this combination? As there is not a sufficient enough amount of publications available at the time of writing this thesis, only an assumption about the behavior can be made. There are two possible outcomes for hybrid conferences. It either can double down on the positive aspects of both physical and virtual conferences, or much worse, the downsides of the two types are multiplying each other. The actual outcome, however, can only be observed over the following years.

# 4 Towards an Evaluation Framework for Virtual Conferences

This section mainly handles the central aspect of this thesis, the journey to a future-proof evaluation framework for virtual conferences in the field of computer science. We build upon the knowledge gained from the conducted analyses in Sections 2 and 3. Furthermore, we try to solve the problems identified using a software engineering-based approach. This solution is achieved with the help of UML diagrams compiled from actors, user groups, actions, and processes identified in the previous sections. Also, requirements for a specific implementation are tried to find out.

## 4.1 First Draft of the Concept at a Glance

After the pre-analysis found in Section 2, we have to first recall our research questions that have been defined in Section 1:

1. RQ1: What is the anatomy of a conference, and how can it be visualized adequately with all parts and features?
2. RQ2: What metrics can be used for evaluating virtual conferences?
3. RQ3: How can we collect and compare the necessary data?

Combining these questions with the knowledge gained from the aforementioned pre-analysis and the current state of conferences, three major topics can be separated that relate to the RQs:

1. Anatomy of a conference and atomic elements of it
2. Criteria that are needed for evaluation and their conversion into comparable metrics
3. Collection of metrics and related challenges

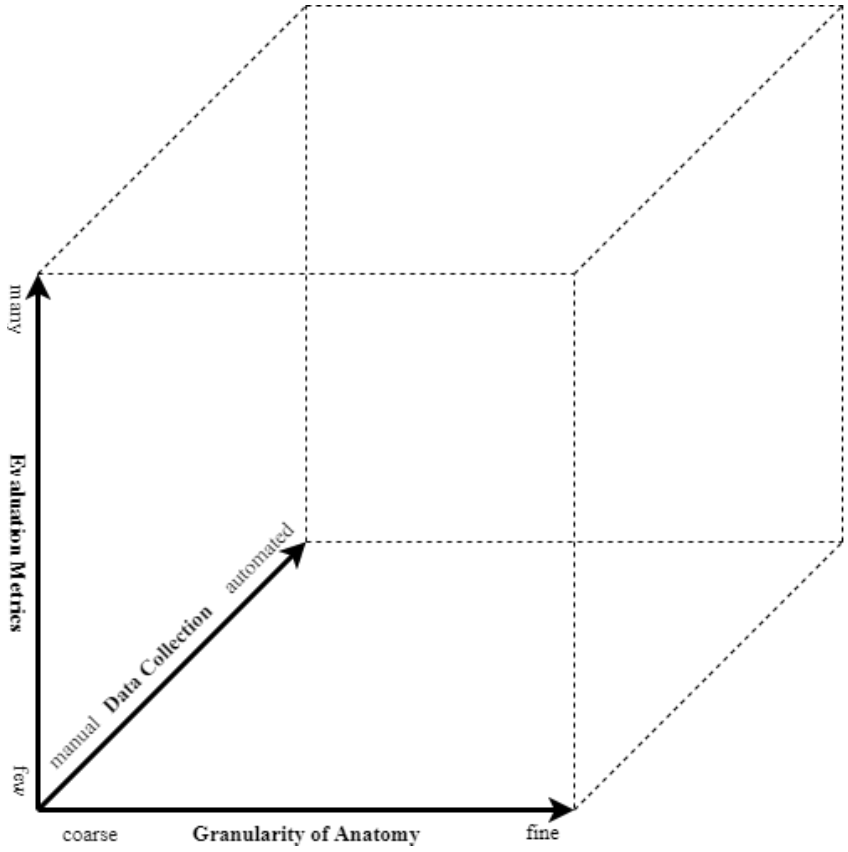



**Figure 4.1:** Cube containing three fundamental dimensions for a virtual conference evaluation system

With these three axes, a cube is spawn, with the appropriate combination for an evaluation framework found within. Figure 4.1 shows the cube with its axis.

The following three dimensions are then relevant for such a system:

- The anatomy describes the core components of a conference and thus the units under test for every evaluation. With this dimension, mainly the granularity of the actions is determined. Examples for such a granularity would be a singular presentation, mapped to a timeslot in a Zoom[1] meeting representing a fine granular item and the whole presentation track over the duration of the conference for a coarse one. For that reason, parts of a conference are examined from a different point of view and displayed in UML use case diagrams to emphasize roles like the session chair and which actions overlap.
- With the evaluation framework, further analysis into the results of (virtual) conferences is executed to show all comparable aspects for these types of events. After that, criteria are then converted into specific evaluation metrics to map the observations in fitting scales. This dimension mainly showcases the number of comparable aspects needed to be representative for all conferences and not be too overloaded. Also, a deeper dive

[1] `https://zoom.us` (accessed: September 2022)

into environmental topics, sustainability, and accessibility is performed, as it is often overlooked with other tools [Bar+21; AI22; Ver22].

- Last but least, this dimension is relatively dependent on the latter one. Therefore, it is essential to look into techniques to collect data that are needed to calculate the chosen metrics for our practical project. However, this is also a complicated topic as legal regulations like the GDPR limit the possible yield of collected data. Also, not everything is automatable to fetch or scrap information from meetings. Some data, like overall satisfaction with the event and its organization, must be collected automatically. Thus, this axis resembles the amount of automation in the data collection process.

## 4.2 Dimension One: Anatomy of a Conference

In this section, various elements of a conference are analyzed to get a comprehensive overview of its anatomy of it. With a deeper understanding of atomic elements, an assessment of granularity on which evaluation is applied can be done. Furthermore, the final goal is to support this decision with a UML use-case diagram, which sets parts of a conference in context with the different conference participants.

### 4.2.1 Key Components

Before such a split into different granularities can be done, a significant question must be answered in advance: What is the anatomy of a conference? Analogous to the human body, even events consist of aspects, parts, sequences, and sub-events that are interwoven similarly to the nerve system. As previously known from Section 2, conferences foster knowledge exchange and enable face-to-face contact between participants to strengthen personal networking. This statement is also supported by Edelheim et al. [Ede+17] as a critical factor for attending conferences. Furthermore, strong networks are essential for exciting discussions, forming cooperation, and initiating new projects. Macek et al. [Mac+12] think that understanding in such contexts is vital for increasing the efficiency and effectiveness of individual networking. They further state that mostly static questionnaire analysis is done, enabling only static behavioral analysis. Therefore, the present dynamic nature of conferences is often not accounted for. However, in light of this analysis, a static structure should be built to create a foundation enriched with dynamic elements.
First of all, which scopes are found inside an academic conference? In the following paragraphs, Learning from Academic Conferences by Celia Popovic [Pop18] will be the central guiding literature, also supplied with additional papers. Although the book focuses on learning, it helps to understand the different facets of such events more. Moreover, due to their

age, conferences are mainly described as real-life physical events. This status quo is a good foundation, as it makes reflection possible on what to exclude for virtual events. An excellent first categorization level can be found when referencing the table of contents. Popovic divides the lessons learned into three points of view:

- Attending a conference
- Presenting at a conference
- Organizing a conference

As a preview to better visualize the processes, refer to the completed UML diagram at the end of this section, see Figure 4.2.

#### Conferences as an Attendant

As seen in this short list, being a participant alone in a conference is not always the same. On the one hand, instead of submitting a paper and hoping for a note of acceptance, attending an event as a listener is possible without hurdles. On the other hand, being a presenter or author at the conference most likely guarantees you an active role in at least one talk rather than being a passive listener.
There are some things to consider before attending a conference, beginning with an attendee's perspective. The participant has to plan the ultimate goal of participating in the event. Remembering the schema proposed by Sanders et al. [San+20], there are different motivational aspects. These motives coincide with the key reasons Popovic stated in the book [Pop18]. In addition to a general insight into other researchers' work, networking, and building their reputation, she also mentioned making time and space to reflect on practice. This reflection is critical because, in today's times, there is much work to do, so often, the process of reflection is neglected. The insight into others' work is also an essential aspect of learning, supported by Reinhard et al. [RSP21]. After this first goal definition, almost as important as the motive is securing funding for the attendance fees, as these entry tickets are often not cheap, careful planning is a must. Participation at a conference should be done at least months before the event itself. Furthermore, possible events should be sorted by the best personal return on investment. However, booking accommodation and traveling is another crucial and costly factor to consider in advance, especially in physical editions. It should also be noted that some universities might need travel proof for travel expense refunds, so keeping all documents is vital. Another factor that suffered from the latest developments of the COVID-19 pandemic is social interaction [SFS19]. Popovic [Pop18] states that meeting other researchers sparks new ideas and helps to learn even more about a topic apart from sessions. Therefore she suggests identifying people to meet by marking their names on the attendant lists. Also, if appropriate, it might be an icebreaker to email them beforehand to establish the first contact regarding a

shared research interest, for instance.

A conference consists of a strictly planned program when remembering Section 2, often divided into multiple tracks. These three to four parallel series of sessions often have a set focus, for example, "Information Systems in e-Government." While switching tracks between sessions is the easiest way, switching during one is a problem. While it might be a simple task in a virtual edition, Popovic [Pop18] states that it may or may not be possible to switch rooms in a physical conference smoothly. Depending on the schedules of session chairs, vital information of a presentation can be missed during the room switch. Therefore she recommends planning and deciding on what to attend in advance. In addition to the last point, it is a fact that sessions are often a particular focus in the vacuum of the track, and contained presentations are grouped for a particular reason. To get the most out of the blocks, the attendee should take notes for future reference and, most importantly, try to listen to the session as a whole. In an optimal case, this should be done without phone distractions.
Another important aspect is the task of networking. There are many actions to do to achieve the best possible experience:

- Checking the conference delegate list
  The delegate list is one of the single most important documents of the whole event. This list allows for a quick scan of who is there and which persons are already known.

- Wearing the conference badge
  This point supports the simplification of introductions and helps identify researchers that were marked beforehand.

- Exchanging business cards
  Handing out business cards helps establish discussion even after the conference. Adding information about the conversation is also beneficial to enable more accessible follow-up meetings.

- Participation in networking events and the conference dinner
  Most conferences offer designated areas for networking opportunities, and they are often used during the session breaks. Also, the benefits of attending a conference dinner outweigh the cost of establishing new contacts, especially when attending is optional. It also allows one to have interesting in-depth conversations with others at the table.

For the context of the content of a conference, Popovic [Pop18] recommends reflecting on the heard presentations and not blindly implementing every new idea that was introduced in the sessions. Therefore notes of what stood out should be taken, and a summary of the key takeaways should be written. Also, reviewing all received business cards is essential to plan whom to follow up with and whom not. From a participant's point of view, completing any evaluation process or feedback around the organizers' offer is very beneficial, as it might

improve future conference editions.

In the aftermath of a conference, Popovic doubles down on the topic of taking notes as the essential action here is to revise the information input of the whole event. Questions like "What were the main messages?" or "What stood out as significant ideas?" are fundamental at this point. As researchers mainly depend on an institute, this information should optimally be shared with colleagues. Therefore a short update in the form of a blog post or a knowledge-sharing session is very beneficial in this context. Also, things about the conference, how the experience was and if it can be recommended should be included. An important factor is also the follow-up with contacts established during the event. Popovic [Pop18] recommends reaching out with further questions via email or inviting them to specific events. Also, more traditional methods like postcards or contact via social media might be a good way. If the recipient does not respond in a few weeks, one reminder should be considered, but not more. Furthermore, last but not least, a conference is set before a conference. It should be considered if repeated participation at the same conference brings additional gain or if being part of other conferences might be better. Thus, planning the event calendar is not a momentous sprint but a continuous effort.

### Conferences as a Presenter

Already discussed in the introduction, every journey as a presenter at a conference begins with the *Call for Papers*. In those, conference organizers publish a document or text on their website with all sorts of general conditions: deadlines, submission formats, and topics for tracks. A large index of examples can be found on the website "Call For Papers"[2]. Regarding the topic of paper formatting, formats from Springer, like Lecture Notes in Computer Science (LNCS), are a popular choice for conferences[3] The chapter in Popovic's book [Pop18] is more orientated on a proposal and presentation type of conference structure. However, some points apply to a format where papers are also submitted. First of all, the paper content. It is recommended to structure the paper in advance to craft headings and annotate them with meta-information like the word- or page limits. This structure supports adhering to the submission criteria of the event. On that note, it is helpful to prepare demo material in case something is presentable as part of the research topic, like a tool. Also, as noted earlier, there are deadlines, so keeping track of them is an important action item. This restriction is especially the case when applying to a keynote, poster presentation, or panel, as there might be other deadlines to consider. However, larger national or international events often have large gaps between the submission deadline and the conference itself, like half a year. The

[2] `https://www.call4paper.com` (accessed: September 2022)

[3] `https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines` (accessed: September 2022)

track topics are always decided in advance so that researchers can submit their work directly for a track. However, depending on the conference, the theme might be less concise than expected, thus making planning harder. For example, a paper regarding the performance of NoSQL databases is suitable for "Database performance engineering" but might also be for "Information Systems and Knowledge Processing." Not to forget, there is also the format of different sessions. For instance, normal paper presentations have another submission format than a workshop paper. Also, the presentation content and time vary strongly. The first one is mostly done within 15-30 minutes and comprises the highlights of the research work. The latter is mostly performed in up to one and a half hours. In an optimal case, it engages the audience to participate in the topic. Even before submitting, it is also helpful to think about the presentation. Are the results better to present alone, or should there be multiple speakers? Although this depends on the conference type and the session format, based on experiences from DEXA and IIWAS, most papers are presented by only one author. It is also worth noting that speakers are mostly students or "lower" rank researchers, with professors and similar personalities letting their peers gain experience in conferences. Remarks on the academic writing process are omitted for this thesis.
Nevertheless, what to consider after the research work is done and a paper is finished and submitted? As an optional step, there may be some critical feedback from reviewers that should be adapted before the work is finally accepted or something to point out in hindsight of a presentation. Also, feedback has to be worked into the paper and submitted again for a so-called "camera-ready" document, which represents a final submission version and is the base of publication.

During the conference, the actions to take are similar to the role of a simple attendant to the event. However, the main difference is that it needs more planning and training done in advance. According to Popovic [Pop18], the presentation should stay true to the submitted document, not invent new content entirely. Especially for a paper talk, it should represent the main findings and inventions in a concise and understanding way, so the audience can follow the topic and is not bored by it. As for the typical case of a session with three to five talks, the time slot is not that extensive for such content. Therefore it is best practice to plan for a talk with a little less time than given. This extra time is especially beneficial when staying calm; with a calm appearance, it is proven to speak a bit slower than a nervous one. Also, this gives room to play and strike out more detailed explanations. In particular, when the schedule is tight, and the session chair asks to shorten the presentation by a few minutes, it helps keep the talk flow. This plan B also allows audience participation in the form of questions. Furthermore, it should be noted that everything from a technical perspective should be tested thoroughly. Respecting Murphy's Law [Blo78], it is recommended to have the presentation on multiple devices and, if possible, to test the functionality at the conference venue. After the talk, Popovic [Pop18] recommends allowing for follow-ups with the audience. Preferred

ways are in physical meetups to show some presence after the session to network a little bit or in virtual editions to leave behind contact information in the chat, for example. She advises thanking the participants for their attention as they might have chosen the paper presentation from all other options. Remembering the last section, they are even explicitly interested in the talk in particular.

In the aftermath of the event, there is only the possibility of further feedback on the topic to add. Also, there is almost always some sort of prize for the best publication at the conference closing ceremony awarded. Within that, the whole reviewer committee ranks the available publications from multiple standpoints and chooses the objectively best one. Finally, there are conference proceedings, where all papers, revised with feedback from the event and if the authors want to, are published in one book.

#### Conferences as an Organizer

In contrast to the points of view, organizing a conference brings a different set of action items with it than being an active or passive participant. First of all, before even thinking about organizing the event itself, consider the reasons why researchers want to attend this conference. The information in Section 3.1 is helpful guidance in understanding this behavior more. Also, Popovic [Pop18] tries to warn about the tips she gave to attendants. As structuring the conference towards taking notes, seeking contact with peers in the same field of interest, and discussing the latest developments during breaks is very dangerous and can lead to lower overall satisfaction with the event. They recommend that the better event experiences are related to an appropriate time and space, also seen in Section 3 as well as with the support of interactivity and active engagement.
Timing is a crucial part of the planning structure when speaking of the former. The time zone is essential, and the time of the year is vital to not clash with a majority of conferences. However, this point must be assessed within the community and generally offers no one-fits-all solution. Another essential factor of success is to have a well-organized team that shares a clear vision of the event to be organized. Therefore team members offer a wealth of experience, interests, and, most importantly, a fair share of time, which can be invested in planning. Popovic [Pop18] suggests one person working on a meta-level as well. This position is responsible for keeping track of the planning progress and keeping the team informed about the tasks. Furthermore, the success of a conference is also carried out by volunteers, often made up of students and facility members. Primarily to deliver a professional and well-organized picture to the participants, it is recommended that volunteers wear the same T-shirt or can be easily identified as staff.
Aside from the helping hands that are crucial for planning, sustainability is another vast topic, which gained popularity in the last years [PMR19]. On the one hand, what is included

in a conference bag should be considered. Does it have to be a unique one? Popovic [Pop18] suggests asking the local bookstore of the institution if they have free ones instead of creating a new one. Also, regarding contained "swag," all of the small things that are free giveaways associated with the event and supporting institutions, consider making them collectible on a front desk instead of including them all in the bags, where a substantial number is being thrown away. It can also be questioned if "swag" is needed. On the other hand, one of the most significant hurdles in event sustainability is disposable cups, dishes, and cutlery. If possible, it is recommended to avoid them at all costs. Otherwise, one can try to purchase recyclable or compostable materials. Also, an engaging way is to encourage researchers to bring their mugs for the sake of the environment.
Last but not least, in terms of pre-event planning, try to work towards goals and the general focus of the event. If it should be a more networking-orientated conference, make breaks longer or organize sub-events that foster discussion and socialization, like town hall meetings and cocktail evenings. Also, feedback from the volunteers and the organization team is greatly appreciated during the organization phase. Furthermore, personal conversation and showing gratitude for the work so far is necessary to keep morale and momentum up for the hard part: the work during the conference.

On the event days, there is, first of all, everything to handle regarding the greeting process and simplifying networking opportunities. Popovic [Pop18] suggests handing out badges containing the names of the participants that may help with identifying each other. Also, if the budget suffices, adding ribbons that highlight the status of the person, like "first-timer" or "reviewer," may help to break the ice of a conversation.
The conference team should be instructed to work on different tasks on a day-to-day basis, comprised of an array of simple to larger jobs. This partitioning helps immensely to stay focused if something out of the blue might happen. Remembering Murphy's Law [Blo78], unexpected events may also occur for the organizing team. The first step in managing this situation is to not reflect on and try to fix structures that were or were not introduced during the conference's planning stage. This strategy is mainly the case since going back in time is impossible. Instead, the only thing that is possible is to do is to fix problems with the best effort as they arise. However, taking notes on all events that went wrong is recommended to prepare for future editions of the conference. The results could also be published as lessons learned papers to disseminate knowledge about holding more successful virtual conferences. Lastly, organizers should also enjoy the sessions in the meantime, as far as possible.

In the post-conference phase, the main goal is to revise the main event and collect data. Data can be used in two ways: (i) collect information about "the good, the bad, and the ugly" to form a report for the administrative team. Furthermore, (ii) retrieve all sorts of photos, video, and text material to make, i.e., the conference website, a record to relive the event. Popovic [Pop18]

refers to the Professional and Organizational Development Network[4] (POD) as a repository for sharing slides and handouts. Also, the conference proceedings mentioned above are a valid option to collect the contributed work of the scientific community to the event. However, the most important thing is to reflect on the past edition and collect ideas and improvements for the next conference. This goal can be achieved through evaluation methods, as seen in Section 2 or discussion sessions at the end of the main days.

### 4.2.2 Granularity of Analysis

The last section mainly handled the topic of identifying the critical components of a physical conference and how to handle them properly from different points of view. With the main focus on virtual conferences for this thesis, this information is sufficient for a general overview of the anatomy of a conference but not for a thorough evaluation. Therefore, the analysis comprises two steps to find the right granularity for that task: (i) distinguish physical-only activities from ones that are possible in a virtual conference, and (ii) separate valuable elements of virtual-only events. In contrast to Popovic [Pop18], the central part of this section is compared to "Realizing an online conference" by Busse and Kleiber [BK20].

#### Core Categorization

There were two categories identified when beginning with the actors in a conference:

- Participant
- Organizational Staff

On the one hand, a participant can be split into an attendant and a presenter in an overlapping specialization. A presenter could either be an author or an external speaker. This distinction is needed because an author is an attendant outside of the paper talk. The same principle applies to an external keynote speaker if he or she attends a session. On the other hand, the administrative staff has many more specializations. From steering committee and session chair to technical host and student, for the volunteer is everything possible. For the sake of simplicity, the primary organizational staff, mainly consisting of professors and conference personnel, is therefore grouped as the main staff. Session chairs and reviewers are also grouped into the latter term, and all other staff is simply called volunteers.
As for the Conference itself, four distinct categories of action items could be identified:

- Publication Process

[4] `https://podnetwork.org/resources/wikipodia/` (accessed: September 2022)

- Organizational Affairs
- Conference Tracks
- Networking

First, every participant in a conference begins with a submitted paper to a platform for managing to do so. See EasyChair[5] for that regard. After the call for papers is finished, reviewers will check the submissions and give their collective feedback. Depending on their decision, which can range to several grades, the work is accepted or rejected. Toivanen [TS22] experienced seven different grades for rating acceptance at the European Association for Research in Industrial Economics (EARIE) annual conference, ranging from A as a definite acceptance down to F as a definite reject. For simplicity of modeling, this results in three different outcomes, linked with their translation to the scale by EARIE:

- Acceptance (A - B+)
- Adapt to Feedback (B)
- Rejection (C - F)

When minor areas need to be rewritten, some reviewers allow them to be revised to change their final rating to an acceptance. However, in most cases, especially with rejections, detailed feedback is given with the motivation to give additional input for a more polished version at another conference. It has to be noted that the review process can be highly subjective. Therefore, multiple reviewers have to be employed per submission. On that topic, Toivanen [TS22] states that out of the 2,261 submissions, even six percent were rated with a high disagreement (one accept, multiple rejections) between critics.
In contrast to adapting to feedback, another type of post-submission paperwork can be seen there. In light of publication for proceedings respectively the conference itself a camera-ready[6] submission is needed. The term originated from earlier times when the paper was sent to a printer to be photographed for offset printing. Although this is not the case anymore, the definition stayed for submissions that are included in the final conference proceedings without any further revision, thus printed "as is".

The second area is more on the organizer's side of the spectrum and primarily contains administrative work. First of all, all first-step reservations and such can be grouped into a "General Affairs" section. This group includes the initial setup of the conference schedule, handling mail requests, and booking the event venue, including all other necessary preparations to book on-site. For this reason, a secretary or personnel in general responsible for that kind of work is especially helpful. For the schedule, the steering committee has to be present

[5] `https://easychair.org` (accessed: September 2022)
[6] `https://www.enago.com/academy/what-is-a-camera-ready-copy/` (accessed: September 2022)

as well. Lastly, in that area, an important aspect is public relations, including notifications and newsletters for participants, but also in the field of social media. Especially in the current times, having a well-structured social media presence is beneficial for newer and existing conferences to establish them as essential brands and to gain more reach in the scientific community. Mason et al. [MNM21] further see that customers and communities after COVID-19 are also shifted towards a more online customer behavior activity.
On the days of a conference, aside from the program and contained sub-events and tracks, there are also ceremonies. The content, consisting of the opening and closing, originates from the organizational and administrative spectrum. Additionally, in the time slot of the closing, also the award ceremonies occur. Most notably, there is the "Best Paper Award[7]", crowning the best submission of the conference out of all submitted works of the conference. The steering committee and the reviewers do the selection process. However, there should be a separate chair for selection by ACM guidelines. Originating from personal experiences from DEXA and iiWAS, there are other awards, for example, for the best student contribution to the event.

The conference tracks are their own subsection when moving on to the main stage of program points. The primary element in this section is the assortment of parallel tracks that logically consist of multiple tracks held simultaneously. In most conferences, these are used for thematic groupings, like a larger topic such as "e-Government" or sub-conferences and workshops. This structure is known from, i.e., DEXA. In these constructs, organizational support from almost all volunteers is needed, as participants and speakers should experience the tracks as smoothly as possible. As introduced before, tracks consist of multiple sessions containing several presentations, mostly three to five [Pop18]. Apart from paper talks, there are also keynotes and panel discussions possible. However, a mix of different types of talks is only possible if there are breaks. This mix is needed to guarantee audience focus and foster participation. Engagement, in general, is an important result of talks, as they provoke discussion and inspire new ideas. This thesis divides audience participation into questions and remarks to show the ratio between them. Remarks in particular also include similarly treated actions like comments and opinions[8].

Last but not least, there is the aspect of networking. As introduced in Section 3.1, networking is one of the main motivations for researchers to attend conferences. Therefore it should be included in this categorization. The main category, which can be used for grouping, is called "Networking Possibilities" and contains smaller opportunities that simply happen throughout the event. These include coffee breaks as a logical consequence of the session structure and hallway chats as informal possibilities to break the ice. Furthermore, the organizer team

[7] `https://www.acm.org/publications/policies/guidelines-best-paper-award` (accessed: September 2022)
[8] `https://grammar.collinsdictionary.com/english-usage/what-is-the-difference-between-comment-mention-and-remark` (accessed: September 2022)

naturally plans several events to foster networking. A staple is the conference dinner, as an opportunity to thank and celebrate the administrative staff for their outstanding work around the event [Pop18]. In a more seldom fashion, also cocktail parties are held to gain distance from the strict and formal central part of the conference. Their primary purpose is to give the participants and staff a chance to meet up and socially interact in an informal way [RMA18].

#### Relevance for Virtual Conferences

As stated in the introductory paragraph for this chapter, not every component in this categorization is usable for every type of conference. Therefore, this section mainly focuses on the topic of filtering out parts that are irrelevant for virtual conferences. Therefore, the four categories are revisited in the same order and compared against the literature regarding virtual conferences.

Regarding the publication process, there is nothing that is affected by a virtual edition of such an event. Thus nothing has to be scraped from here.
Similarly, organizational affairs are not much changed either. One of the most notable changes for the personnel is not to be responsible for booking an event venue and, therefore, not having to organize the surroundings. According to Busse and Kleiber [BK20], ceremonies like the opening and the closing are gaining enormous relevance because they should be held in a live format in both synchronous and asynchronous virtual conferences. For many participants, this is one of the only chances to see and get to know other participants and, more importantly, the organizational team. Furthermore, it guides new participants on what to expect in the sessions.

Regarding the conference tracks, there is also no change to be expected, as the system for parallel tracks is mainly kept, except for asynchronous conferences, as explained in Section 3. In such cases, the concept of (parallel) tracks is omitted. It has to be noted that there should always be good moderation in place. Moderation is vital for participant engagement. Busse and Kleiber [BK20] believe that moderation and netiquette improve the overall satisfaction of the audience with the session. Examples of well-organized netiquette could be prefixing questions with "Question:" or raising the hand in Zoom before asking. However, in their questionnaire regarding their conference at the International Computer Archive of Modern and Medieval English (ICAME), they received feedback where the process of taking turns in a discussion was seen as a flaw.

The most considerable reduction of possibilities can be seen in networking. Nearly all of the informal and non-planned networking opportunities are erased here. This whole section could be crossed out when seeing the events as only physically possible happenings. However, with the rising efforts of the community, as seen in Section 3 and with virtual events, like cocktail parties or virtual city tours [RMA18; BK20], only the conference dinner is affected. Although their effectiveness is something to be desired, the potential alone lets this subsection survive the compression.

### 4.2.3 Complete Schema

Based on the information from before, a three-level UML use case diagram schema is proposed:

- General overview displays possible actions for standard physical conferences
- Virtually possible actions
- Virtually possible action items that can be evaluated

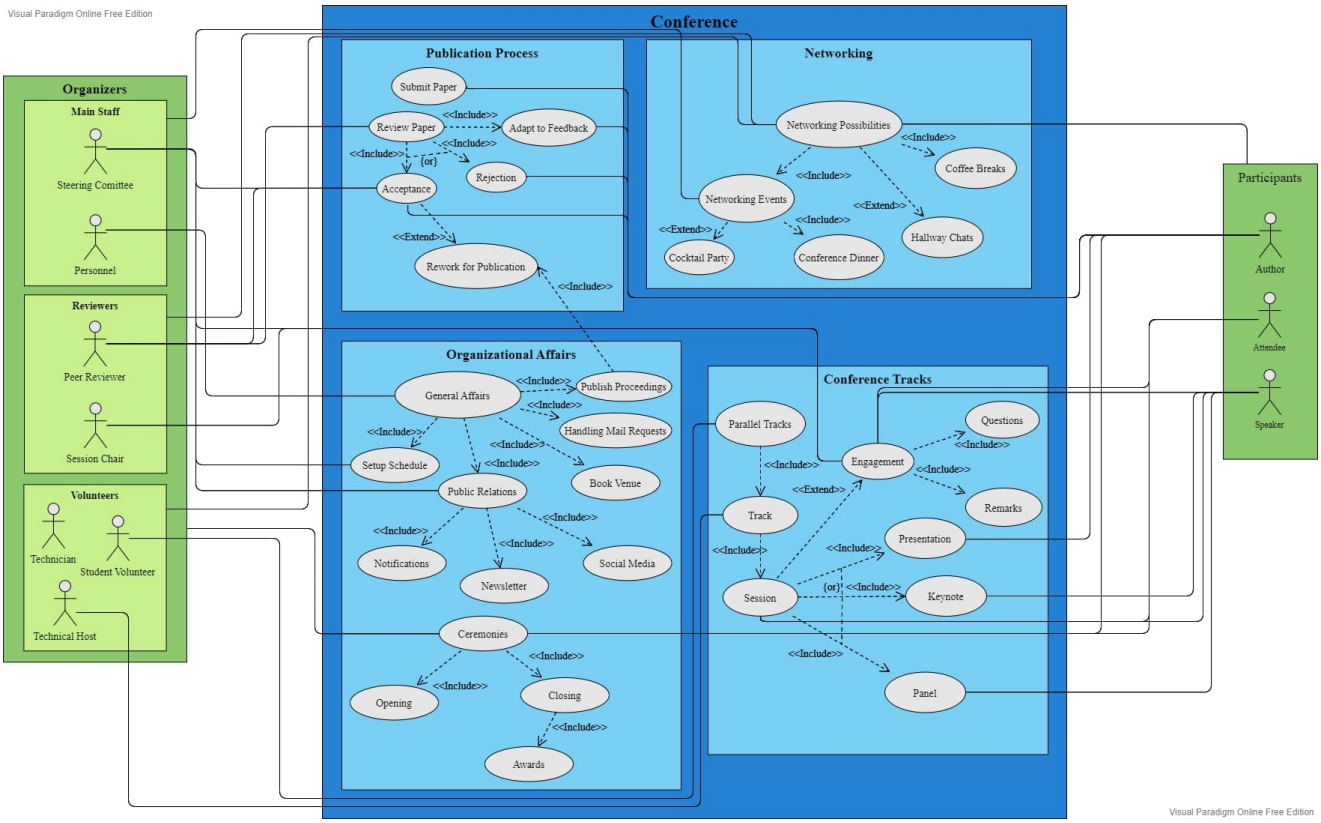



**Figure 4.2:** General overview of a conference as a UML use case diagram

Following from Figure 4.2, some actors are grouped via the categorization from Section 4.2.2. Associations that affect all actors of that group start from the group block. Thus, readability in the diagram is increased. However, this style might seem odd at first, as the technician of the volunteers might look like they have no link associated.
The diagram also has to be adapted by seeing the limitations of virtual conferences. As seen in Figure 4.3, action items are erased, and if needed, some associations have been rerouted. Also, the concept of tracks has been marked because of the restrictions on asynchronous events.

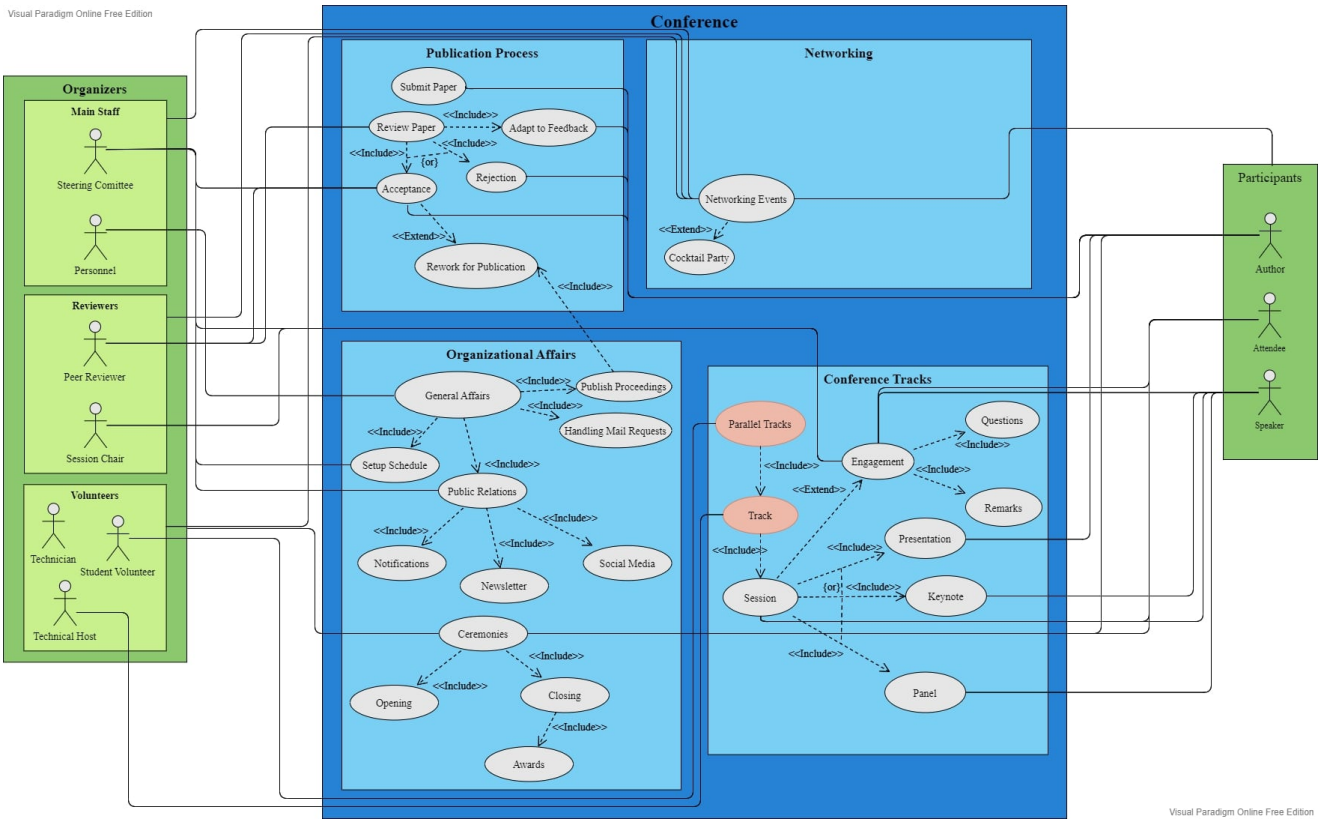



**Figure 4.3:** Adapted use case diagram from Figure 4.2 for virtual conferences

Finally, the current use case diagram should be cleansed for evaluation in light of the evaluation framework. Therefore, only a few categories that provide valuable information should be left afterward.
All action items regarding the review process were omitted, as personal feedback is irrelevant for conference-wide evaluation. Furthermore, the rejection is also erased. This deletion is because acceptance stands for the closely related *Acceptance Rate*[9], which is an essential factor when comparing (virtual) conferences. However, it should not be the primary and lone factor.

[9] `https://people.engr.tamu.edu/guofei/sec_conf_stat.htm` (accessed: September 2022)

In terms of organizational affairs, the general affairs category was also kept to remind the organizers of all economic metrics based on sustainability and Eco-friendliness. Further, the published proceedings can also be used to keep track of the ongoing research impact of the submitted papers. The newsletter and social media can measure overall engagement with the community and the reach of announcements. The ceremonies are also kept to help determine the number of participants in the sessions, as the opening and closing are attended mainly by the majority of participants.

As for the conference tracks, this section was reduced to the absolute minimum because more information cannot be gained from the single sub-parts of a session, as most such presentations are grouped into one special virtual meeting. Therefore, additionally to the session itself, only the engagement of the participants remained as well. From those, interesting measurements such as minimum and maximum of participants, all statistical aggregations, and more interesting user-based interaction ratios can be taken.
Last but not least, the networking area has not been changed from the last iteration. The results of this process can be finally seen in Figure 4.4. This structure now represents a medium granularity since there are still fine-grained components in this diagram, especially in the area of organizational affairs, with newsletters, for example. However, on the one hand, other elements, like the presentation or questions and remarks below engagement, are gone. On the other hand, the coarse components are also removed. This upper and lower bound towards components leaves mostly only medium granular items in this use case diagram.

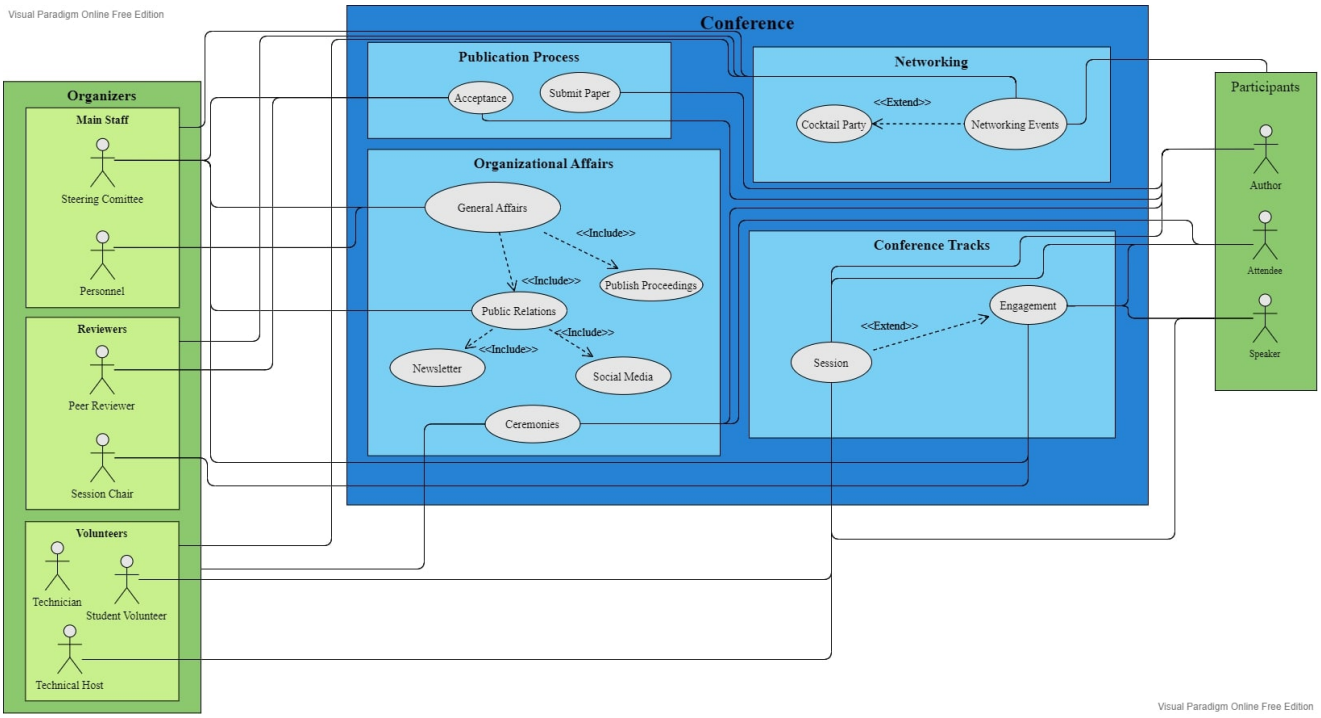



**Figure 4.4:** Adapted use case diagram from Figure 4.3 that shows virtual actions that can be evaluated

## 4.3 Dimension Two: Evaluation Criteria and Metrics

In this section, we perform the next step of the analysis by extracting common reportings and themes from our factual base and relevant literature and condensing them into evaluation criteria, thus laying the groundwork for answering RQ2 and proposing an evaluation criteria fingerprint that captures the individual characteristics of a virtual conference in the next section. The work done in the previous dimension supports this process by steering the research into a pre-filtered focal area.

### 4.3.1 Purpose

The primary purpose of these evaluation criteria is to aid organizers in choosing the right framework conditions for their conference. We put forward the implementation of an extensible reference framework for evaluating virtual conferences, which can be used for evaluating conferences based on selected metrics. The evaluation criteria serve as a guideline in which areas a conference should be monitored. However, as the reference framework is intended to be compatible with any type of academic conference in the field of computer science, the evaluation criteria and suggested sample metrics just serve as a starting point and should be tailored by an expert to the specific conference. The reasoning behind the use of evaluation criteria is to allow for comparison, identify improvement potential, and help describe lessons

learned. Ultimately, taking these considerations into account will lead to better iterations of the conference in the future.

### 4.3.2 Evaluation Criterion 1 and 2: Accessibility and Usability

Accessibility touches on all aspects of a conference, e.g., location, website, submission and review process of papers, and social events. Whether or not a conference was designed with accessibility can impact and even prevent the participation of people with disabilities. For this reason, we argue that accessibility is not only a critical evaluation criterion but needs to be kept in mind from the very start and evaluated well in advance before the start of a conference. We define accessibility following the W3C Web Accessibility Initiative [Hen] in the following way:

> *Accessibility means that websites, tools, and technologies of a conference are designed and developed in a way so that people with disabilities can use them to contribute, interact, navigate, perceive, and understand.*

However, people without disabilities also benefit from accessibility measures geared towards people with disabilities [Hen]. For instance, people using mobile phones, people having a slow internet connection, or aging people could benefit from transcripts. In terms of measuring this criterion, we recommend the Deque Web Accessibility Checklist of Deque University [Uni], which is based on the WCAG 2.1 AA (Web Content Accessibility Guidelines, a W3C recommendation).

We can therefore define the following metrics for accessibility based on the principles of WCAG 2.1 AA[10]:

- Perceivable: All information and components of the user interface must be able to be registered by the user by at least one of their senses. This principle encompasses twenty success sub-criteria.
- Operable: Users must be able to use components of the user interface and navigation successfully. Users must not be barred from access by interactions they cannot perform. This principle comprises seventeen sub-criteria.
- Understandable: Information and the user interface must be presented in a way that can be successfully cognitively processed by the user. This principle encompasses ten sub-criteria.

[10] `https://www.w3.org/TR/UNDERSTANDING-WCAG20/intro.html` (accessed: September 2022)

- Robust: Content must be reliable enough to be used by users and assistive technologies. This principle consists of three sub-criteria.

All four metrics consist of a total number of reachable points corresponding to the number of sub-criteria belonging to a principle. A point is obtained for every met sub-criterion. In total, fifty points can be obtained. As each principle should be fulfilled as completely as possible, a conference cannot be counted as accessible if even one of the principles fails. Therefore, we suggest including a threshold of a fixed value, e.g., 80 % of the total number of points have to be reached to meet the requirements for an accessible conference.

The WCAG EM Report Tool[11] provides the functionality to fill out a questionnaire for all four principles and their according success criteria and to generate and download a report in HTML or JSON. The whole report or the total scores on the four principles can be considered for the evaluation of a conference.

Tied closely to accessibility, we also need to consider usability. In reference to Alonso-Virgos et al. [Alo+20], we define usability in the following way:

> *Usability is the degree of effortlessness, effectiveness, and efficiency of the use of a conference in terms of its contents, tools, and websites.*

Not only disabled people profit from increased usability of the web conferencing tool or website. Instead, all participants can have a more rewarding experience during a conference when the effectiveness, efficiency, and effortlessness of using a website are well-designed. Alonso-Virgos et al. [Alo+20] recommend user-centric evaluations, which can include tests, interviews, and physiological measurements since the effortlessness of use of a website or a tool is best discovered by users with little training in using it. Additionally, usability inspections by experts following usability guidelines or checklists could be utilized.

Another critical aspect of accessibility in combination with usability is the users' acceptance of the chosen tools. From their evaluation of the miTAS Project, Heitplatz et al. [Hei+20] learned that acceptance of technology is a process that can be divided into three phases. First of all, an attitude phase, in which the users are confronted with new technology for the first time and preliminary evaluate whether it is beneficial to use it based on the available information about the technology (e.g., introductory texts, onboarding slides, and instructions). Then, the action phase ensues, in which the users engage with the new technology and try it out more intensely. Finally, the usage phase begins when the new technology is deemed valuable, and the acceptance level has risen. The new technology is used for specific tasks and reassessed

[11] `https://www.w3.org/WAI/eval/report-tool` (accessed: September 2022)

during this trial. Additionally, Heitplatz et al. [Hei+20] identified four factors that contributed primarily to the acceptance of technology:

- Perceived usefulness
- Ease of use
- Job relevance
- Enjoyment of use

These factors are linked to the effortlessness, effectiveness, and efficiency aspects of usability that we have identified, with the sole exception of job relevance. We reinterpret this factor as task relevance to signify the perceived relevance of the technology for taking part in a conference. E.g., if the tool is adequate and related to the tasks that must be accomplished to connect with other conference attendants, give and follow presentations, and interact with provided materials. With this slight adaptation, we deem the four factors suitable for evaluating accessibility and usability since they could be measured with metrics based on the Likert scale in user-centric tests.

We define the metrics for usability, therefore, as the following:

- Perceived usefulness: How efficient and effective is this specific technology in satisfying the informational needs of a user?
- Ease of use: How effortless is the use of this technology for the user?
- Task relevance: How relevant is this specific technology to the current task the user needs to perform, e.g., present a talk, connect with other participants, or submit a paper?
- Enjoyment of use: How accommodating and attractive is this technology to the user, and how likely will the user continue to use it?

These metrics can be taken as the basis for usability tests and can each be rated by the users who conduct the test on a five-point Likert scale.

### 4.3.3 Evaluation Criterion 3: Technical Background

This evaluation criterion consists of different aspects that relate to a conference's utilized tools and website. As technical setups can be very diverse, this evaluation criterion should be tailored to the specific conference. We define the technical background criterion in the following way:

*The technical background criterion comprises measurements regarding all utilized frameworks, tools, websites, and other parts necessary for hosting, navigating, and interacting with a conference.*

The list of sub-criteria can be appended as seen fit, but we suggest starting with the following ones:

1. Amount of tools: In this sub-criterion, the use of tools should be considered and evaluated. We recommend a minimum number of tools that can still achieve the necessary functionality.
2. Integration: The ease of integrating optional tools into the conferencing setup, e.g., connection to social media.
3. Reliability: The stability and availability of the services (up-time and stable internet connection).
4. Audio quality: Essential for a conference, even more, important than the video quality [TB11]
5. Common visual platform [TB11]: Even if other tools are utilized, a common visual platform should be used. Otherwise, the number of interactions between participants will be reduced.
6. Process pipelines: Evaluate if efficient submission and review process pipelines are in place.

The minimum number of tools for a virtual conference includes a video conferencing tool, a website, and a review/submission system. For physical conferences, this aspect is negligible, whereas a hybrid conference would also require a tool that allows for interaction between remote and physical attendees.
Concerning the aspect of integration, it is beneficial for virtual conferences to be able to add more functionality as needed. A high amount of flexibility helps mitigate problems that might occur in the short term. Therefore, APIs and available extensions should be checked. Since this aspect requires technical knowledge, there is no easy way to evaluate this. However, it should be documented and at least put in a verdict on whether the tools are compatible and if there are any obstacles in the setup.
Reliability of the services can be tested beforehand with stress tests, e.g., with tools such as JMeter and Artillery. In any case, reliability should be >95 % to ensure an excellent conferencing experience without too many interruptions.
Audio quality could be measured with metrics such as delay, distortion, echo, frequency response, noise cancellation, signal-to-noise ratio, and total harmonic distortion. Nevertheless, audio quality cannot be measured with numbers alone. It should also be tested for subjective

quality by test users.
Statistics of the standard visual platform and the process pipelines can be tracked and evaluated automatically, e.g., by utilizing the API of the tool to generate a statistics report.

Regarding the apparent problem of the scarcity of expert reviewers for papers and how to solve their assignment to papers optimally, without conflicts of interest, we can recommend investigating the following approaches:

1. "An automated conflict of interest-based greedy approach for conference paper assignment system" [Pra+20]
2. "ReviewerNet: A visualization platform for the selection of academic reviewers" [Sal+20]

### 4.3.4 Evaluation Criterion 4: Reception

From our factual base and the relevant literature [Ful+20; RSP21; LC20; RMA18; Pat+20], we identified reception as an important evaluation criterion. We define reception in the following way:

> *Reception as an evaluation criterion comprises both the anticipated and actual popularity and quality of a conference.*

By defining reception in such a manner, we achieve that it is both relevant for evaluations done before the start of a conference and for reviewing purposes afterward. Furthermore, we define the following sub-criteria for reception:

1. Interaction quality: The level of interaction quality from an academic and a social perspective.
2. Networking possibilities: The perceived networking possibilities.
3. Attendant dynamics: The interaction dynamics, regarding position or dominance, between members of the same group, e.g., virtual, and members of different groups, e.g., in-person and virtual. Other aspects can also be considered, such as cross-faculty interaction and interaction between early-career researchers and late-career researchers.
4. Participant satisfaction: The amount of satisfaction of conference participants.
5. Technical problems: The amount and frequency of technical problems that attendants of a conference encounter.
6. Feedback possibilities: Which feedback channels or methods are available?

In terms of which measures can be used to improve the reception of a virtual conference, we take a closer look at the lessons Fulcher et al. [Ful+20] learned. In the paper "Broadening Participation in Scientific Conferences during the Era of Social Distancing," Fulcher et al. [Ful+20], the authors describe the outcome of a hybrid workshop that engaged both physically present and virtual participants. In general, they find virtual academic conference assessments to be not well researched. Fulcher et al. [Ful+20] also report that even though they followed existing guidelines for setting up virtual conferences, they encountered problems regarding participation and networking opportunities of the virtual-only participants. However, these issues became apparent in the post-workshop self-assessment, which included participant surveys, platform usage statistics, and informal attendee feedback. The identified deficiencies for virtual participants encompassed a lack of networking across disciplines, a lack of informal networking opportunities, and a lack of communication with in-person participants. A measure that proved successful and led to high interaction levels was the establishment of virtual communities (VC), to which the virtual participants were assigned. A discussion was encouraged by trained moderators for each VC, and a mandatory meetup of all VC members before the actual start of the conference was set up.
Similarly, Song et al. [SRM21] used their tool Minglr, a JavaScript-based open-source web app, to help find conversational partners at a conference. Aside from providing the random encounter aspect that many misses from not being able to attend conferences in person, Minglr also allowed private sessions to emulate a private chat that a small group of people would have at a conference venue.
Similarly, Parncutt et al. [Par+21] suggest to include tools for remote immersive conferencing, such as Third Space[12] and Gather Town[13]. In addition to using these tools, they advise planning specific virtual socializing events, such as the following:

- Focus groups that discuss specific topics and allow participation from anyone interested
- Supervisions that consist of prearranged meetings of mentors and early-career researchers
- Reunions that bring together participants that are already acquainted with each other
- Free sessions, i.e., random meetings for different group sizes

Another approach to facilitate social interactions during virtual conferences is Le et al. [LMO20]'s social virtual environments (SVE) that utilized Mozilla Hubs as a platform for remote participants that could engage in co-watching conference talk live streams. To supplement this approach, they used virtual poster sessions, improving attendees' social connections.
In theory, virtual reality could be a solution to design conferences with more interaction and networking possibilities. However, as Lahlou et al. [Lah+21] state, virtual reality installations

[12] `http://www.th1rdspac3.com/` (accessed: September 2022)
[13] `https://www.gather.town/` (accessed: September 2022

of virtual conferences (IVC) are still suffering from the same issues as 30+ years ago and are still a long way before IVC is more widely used.
Palmer et al. [Pal21] recommend the tool Sococo concerning these suggestions, which consists of an online workspace with voice, video, and screen share functionalities[14]. Another way to increase attendee participation is to include tools such as Crowdcast, Mentimeter, or Slido [Pal21]. These tools can, for example, be employed to anonymously vote on questions, improving the quality of the questions posed to the speaker of a talk and providing early-career researchers with a better chance of their questions being picked.
Hohlfeld et al. [HGD21] designed a pre-conference and a post-conference questionnaire to measure the reception, which they analyzed with scripts.
In conclusion, we recommend first setting up reception goals, arranging for post-conference feedback channels via different methods (e.g., interviews and surveys), and comparing the set goals and the feedback. Furthermore, for enhancing the experience of virtual-only users, SVE, VC, or Minglr can be recommended. If applicable, the tools' collected statistics or even metrics could be reused for evaluating the reception.

### 4.3.5 Evaluation Criterion 5: Economic Aspects

Another essential theme we encountered in our literature research and our factual base is that the traveling time and costs provide a significant barrier to conference participation. Even though virtual conferences may be perceived as lacking in the social interaction department, the number of participants increased due to lower costs [Neu+20; Woo+21]. We define this evaluation criterion in the following way:

> *Economic aspects of a conference include all financial measures and resources related to hosting a conference, spanning from its infrastructure to the housing of its participants.*

To differentiate this criterion more, we define the following sub-criteria:

1. Attendance prices: The cost of attending the conference.
2. Internet costs: The cost of hosting the conference in terms of internet connection. From participants' view: The cost of attending the conference in terms of internet connection.
3. Travelling costs: If applicable, it signifies the traveling costs to the conference venue.
4. Housing: The cost of housing the participants.
5. Catering: The cost for the provided food at the conference or in the hotels.
6. Return of funds for research: The amount of money that can be transferred directly to research.

[14] `https://www.sococo.com/` (accessed: September 2022)

However, it should not be underestimated that poor digital infrastructure can also provide a significant barrier to attending a conference. For example, Ozili et al. [OA20] reported that 30 % of US colleges and universities already had a weak operating performance in digital infrastructure and online education. So it was challenging to deal with the required changes to quickly adapt to the new status quo of disrupted learning under pandemic circumstances. If universities are to host virtual conferences successfully and frequently, the cost of developing the digital infrastructure should also be weighed. This prerequisite, however, is a prerequisite of generally hosting conferences, so we do not include it as a separate criterion in our set of evaluation criteria.
To measure this criterion, we suggest collecting detailed cost data of previous conferences in each of the six sub-criteria and employing metrics such as mean, median, and standard deviation. The evolution of the costs can then be tracked in a dashboard for each conference and easily be compared to previous years. Another possibility is to calculate the price trend over the years since holding the first conference in each of the six sub-criteria.
Cost statements for each setup should be prepared and evaluated to compare different conference setups for the current conference. This step is done by displaying which setup is the most cost-effective and how much cheaper or more expensive another setup is by either displaying the difference in costs in each category in raw numbers or percentages.

### 4.3.6 Evaluation Criterion 6: Ecological Impact and Sustainability

One of the pervasive themes of our factual base and our selected literature [Neu+20; Pie+20; TB11; Wil+21; Woo+21] was the interest in combining the drive to more virtuality with more environment-friendly behavior. Even amidst the coronavirus pandemic, the climate crisis was an important topic. We define this criterion in the following way:

> *The ecological impact and sustainability concern the carbon footprint and general impact on the environment of a conference and its participants.*

It is important to stress that every participant needs to be considered with their contribution to the carbon footprint of a conference, be it due to traveling or by tuning it virtually (e.g., live streaming). We define the following sub-criteria:

1. Carbon footprint: The calculated total carbon footprint of a conference. For the calculation, a Life-Cycle Assessment (LCA) could be used. Alternatively, carbon accounting could be used according to the Greenhouse Gas (GHG) protocol.

2. Sustainability: A ranking of how sustainable the conference is.

3. Green Innovativeness: The number of utilized measures to reduce the carbon footprint is awarded ten innovation points per measure. The amount by which the carbon footprint has been reduced compared to the average of the three previous conferences is awarded

ten innovation points per percentage point of decrease of the total carbon footprint. Green Innovation Levels need to be defined for the specific conference, as conferences can vary significantly in size and setup. Employing ten new green measures may be a considerable effort for a smaller conference but not for a conference of larger dimensions. Therefore, the green innovation levels should be defined by the sustainability chair or a committee of experts once for this specific type of conference and then adhered to in the following years.

Twine et al. [TB11] adds that many organizations seek to reduce employee travel time and travel costs with environmentally friendly IT practices while also reducing CO2 emissions.
Pierce et al. [Pie+20] argued that scientific societies considerably impact climate change since traveling to conferences represents a substantial part of a researcher's CO2 footprint. Pierce et al. [Pie+20] propose that each ACM-sponsored conference should publicly report its carbon footprint and that a surcharge should be imposed on conferences based on this footprint.
Williams et al. [Wil+21] reported that conference attendees of the ACM Symposium on User Interface Software and Technology (UIST) in 2019 and 2020 produced, on average, two tonnes of carbon for air travel. Location, traveling, and catering was identified as major environmentally detrimental factors. However, they also mention memorabilia and single-use plastic badges are unnecessary factors that could easily be remedied. Another aspect to consider is the rebound effect of virtual conferences that could lead to more physical meetings in the future since more connections have been established during virtual conferences. Williams et al. [Wil+21] state that the advantages of hosting a virtual conference are evident from an environmental perspective. However, from a social point of view, collaboration and networking are areas in need of improvement.
Neugebauer et al. [Neu+20] conducted an LCA to identify the most carbon-contributing parts of a conference. The dominant factor proved to be the traveling, but catering, hotel overnight stays, and environmental burdens concerning the conference venue were also relevant factors. Furthermore, Neugebauer et al. [Neu+20] suggest that the social benefits of direct personal and globally-oriented exchange could not be outweighed by environmental savings, which is why future conference planning should relate the sustainability benefits with the detrimental impacts. Virtual meetings were identified as a further measure of environmental optimization. However, the authors did not deem there possible to be a significant switch to virtual conferences in the foreseeable future. It is important to note that the paper was published in 2019, before the coronavirus pandemic outbreak.
Woodruff et al. [Woo+21] point out that despite having a clear benefit in terms of environmental impact, researchers in the medical and scientific community are less likely to submit papers to virtual conferences because the social aspect of knowledge dissemination is considered to be a critical factor. Future research should focus on facilitating this need and finding an appropriate virtual substitute for conferences that can not be held sustainably in a physical location.

Williams et al. [Wil+21] see the challenge of changing the way conferences are held because many decisions on where and how to host a conference happen many years in advance. As conference sustainability chairs, Williams et al. [Wil+21] are often presented with faits accomplis regarding the choice of the conference venue and other critical aspects for hosting a sustainable conference. Therefore, pro-environmental measures should already be part of the policy of conference decision-makers.
A measure they can recommend for every conference is carbon offsetting, e.g., buying carbon credits to balance the conference's and its attendees' carbon emissions. Since the calculation of carbon emissions at a conference is non-trivial, it is essential to utilize metrics and automation for as many sub-areas as possible to reduce the calculation's complexity. Williams et al. [Wil+21] also state that to their knowledge, there is no research available on the comparison of conference formats and their advantages and disadvantages in terms of their environmental and social impact. Bonnet et al. [Bon+22] argue that a hybrid conference offers excellent potential for hosting more sustainable conferences, especially regarding cutting down on the carbon cost of traveling. However, more research needs to be conducted on possibilities to offset a conference's estimated total carbon footprint.
In order to measure this criterion, we define the following metrics:

- Carbon footprint: The total carbon footprint of a conference in kilogram carbon dioxide emissions per day of the conference
- Sustainability: The ranking that the sustainability chair or another expert attributes to the conference is based on three levels: highly sustainable, sustainable, and not sustainable.
- Green Innovativeness: Total number of utilized measures to reduce the carbon footprint compared to previous years.

### 4.3.7 Evaluation Criterion 7: Privacy and Security

We found that privacy and security are often overlooked aspects. Researchers that presented their lessons learned in our factual base pointed out that security was often an afterthought or that the need for more security arose during the hosting of the conference [Roo+20; HH21]. Therefore, we included this evaluation criterion and defined it in the following way:

> *Privacy and security as an evaluation criterion comprise measures to judge the threats to a conference, its security standards, its compliance to media publication guidelines, and the GDPR.*

As sub-criteria, we define the following ones:

1. Threats: Amount of identified threats or risks to the conference. It can be estimated with a cybersecurity standard, e.g., NIST or ISO 27001.
2. Level of security standard: A ranking based on estimations of standards, e.g., NIST or ISO 27001.
3. Compliance with media guidelines: Measures adherence to media guidelines relevant to the conference. Relies on best practices.
4. Compliance to GDPR: Measures adherence to guidelines of the GDPR.

In their empirical study, Ovrelid et al. [OBT21] state that research data is often managed in unsatisfactory ways since the data is stored on PCs and servers, which they find lacking in terms of security mechanisms. They analyzed the research data platform TSD of the University of Oslo, which should safely store sensitive data. For a secure architecture, they suggested building a platform core on which the sensitive data is stored, which should then be secured by two-factor authentication to prevent access from unlicensed persons and encryption and SSL certificates. Access to the hardware must also be limited to as few persons as possible. The resources that should be made available for the users, e.g., secure REST endpoints, are controlled by security mechanisms and serve as intermediaries between the platform core and the dedicated user services, e.g., apps accessing the resources. Another recommendation to create a secure research data platform is the use of DevOps practices for employing policies, auditing, data management, and data life cycle handling.

It is advised to take the amount of fulfilled criteria from the selected standard or guideline as a base to measure this criterion. The metrics that we suggest to collect for this criterion are the following:

- Threats: The total number of threats identified with a cybersecurity standard, e.g., NIST or ISO 27001
- Level of security: The self-assessment score according to the selected standard
- Compliance to media guidelines: Self-assessment based on best practices with three levels: full compliance, partial compliance, and noncompliant.
- Compliance to GDPR: Self-assessment according to guidelines with three levels: full compliance, partial compliance, and noncompliant.

### 4.3.8 Evaluation Criterion 8: Modality

Even though the focus of this master thesis is on virtual conferences, we want to take into account the observed trend towards hybrid conferences. Especially with regard to future work, it is important to note the details of the conference delivery mode in order to make conferences more comparable and to better study whether hybrid conferences enhance or mitigate the advantages and disadvantages of the other forms. We define this final evaluation criterion in the following way:

> *The modality of a conference signifies the main type of the conference, i.e., hybrid, physical or virtual.*

Furthermore, we define the following sub-criteria:

1. Type of conference: Physical or virtual.
2. Hybrid parts: Whether or not the conference has aspects of the opposing type (physical or virtual).
3. Goals: A list of individual goals for this specific conference, e.g., increased cross-faculty interaction and number of paper submissions

It is mainly used to document the choices regarding hosting the conference, comparing and evaluating what has worked, and leave room for improvement. For this reason, a questionnaire that includes selectable types for the conference and the hybrid parts and text fields to document the choices is suitable for this criterion. A metric for this evaluation criterion could be the degree of virtual parts percent in order to be able to compare the modality of previous conferences.

### 4.3.9 Sample Metrics

As a foundation for answering RQ2, which metrics can be used to evaluate virtual conferences, we have created an overview of exemplary metrics in the following section; see Table 4.1. This table is only a representative selection that covers at least one example of each sub-criterion. The selection is based on the evaluation criteria described in the previous sections. The sample metrics are intended to be used in two areas: First, the sample metrics can be used directly in the reference framework, but can also be extended to include arbitrary metrics, as explained in Section 5. Second, the sample metrics are also intended to provide a starting point for evaluating future conferences. They are not meant to completely cover all evaluation criteria in all details, as we aim at a generic use of the reference framework and it is not possible to predict which additional metrics may be relevant for any given conference. Rather, they serve to demonstrate how the evaluation criteria can be translated into metrics as

examples. We will further refine and discuss this table of sample metrics in Section 4.4, when we delve deeper into the collection of data for metrics.

| **Evaluation Criterion** | **Metric** | **Description** |
|---|---|---|
| Accessibility and Usability | Perceivable score | Based on WCAG 2.1 AA principles, twenty sub-criteria, each fulfilled sub-criterion amounts to 1 point, score consists of the total number of reached points, see Section 4.3.2 |
| Accessibility and Usability | Operable score | Based on WCAG 2.1 AA principles, seventeen sub-criteria, each fulfilled sub-criterion amounts to 1 point, score consists of the total number of reached points, see Section 4.3.2 |
| Accessibility and Usability | Understandable score | Based on WCAG 2.1 AA principles, ten sub-criteria, each fulfilled sub-criterion amounts to 1 point, score consists of the total number of reached points, see Section 4.3.2 |
| Accessibility and Usability | Robust score | Based on WCAG 2.1 AA principles, three sub-criteria, each fulfilled sub-criterion amounts to 1 point, score consists of the total number of reached points, see Section 4.3.2 |
| Accessibility and Usability | Total Accessibility Score | Based on the four scores for the WCAG 2.1 AA principles, the sum of all four scores, the maximum number of reachable points is 50, see Section 4.3.2 |
| Accessibility and Usability | Perceived usefulness | Rated in usability tests, in percent or measured with five-point Likert scale, see Section 4.3.2 |
| Accessibility and Usability | Ease of use | Rated in usability tests, in percent or measured with five-point Likert scale, see Section 4.3.2 |
| Accessibility and Usability | Task relevance | Rated in usability tests, in percent or measured with a five-point Likert scale, see Section 4.3.2 |
| Accessibility and Usability | Enjoyment of use | Rated in usability tests, in percent or measured with five-point Likert scale, see Section 4.3.2 |

| Evaluation Criterion | Metric | Description |
|---|---|---|
| Technical Background | Amount of tools | The total number of tools used for the communication channels, the platform, the review, and submission system, and the website of the conference, integer, see Section 4.3.3 |
| Technical Background | Integration | The level of ease of integrating optional tools into the conferencing setup, e.g., connection to social media, select from levels: not flexible, flexible, highly flexible, see Section 4.3.3 |
| Technical Background | Average response time | Mean response time in ms, see Section 4.3.3 |
| Technical Background | Subjective audio quality | Rated audio quality by testers, select level from: not feasible, bad, good, great, perfect, see Section 4.3.3 |
| Technical Background | Jitter | The fluctuation of the delay of received packets in ms, collected by Zoom[15] |
| Technical Background | Latency | Amount of packet delay in ms, collected by Zoom[16] |
| Technical Background | Statistics of the common visual platform | Statistics report of the common visual platform, e.g., Zoom, see Section 4.3.3 |
| Technical Background | Process pipelines | Evaluated efficiency of submission and review process pipelines in per cent, see Section 4.3.3 |
| Reception | Total number of speakers | The total amount of speakers for a conference, integer, see Section 4.3.4 |
| Reception | Total number of talks | The total number of talks for the duration of the conference, integer, see Section 4.3.4 |
| Reception | Total number of participants | The total number of participants for a conference, integer, see Section 4.3.4 |
| Reception | Average number of questions per talk | The mean number of questions asked via chat for all talks of a conference, see Section 4.3.4 |
| Reception | Average number of participants per track per day | Mean number of participants per track and per conference day, float, see Section 4.3.4 |

[15] `https://support.zoom.us/hc/en-us/articles/204654719-Dashboard`, accessed: September 2022
[16] `https://support.zoom.us/hc/en-us/articles/204654719-Dashboard`, accessed: September 2022

| Evaluation Criterion | Metric | Description |
|---|---|---|
| Reception | Maximum number of participants per track per day | Maximum number of participants per track and per conference day, float, see Section 4.3.4 |
| Reception | Minimum number of participants per track per day | Minimum number of participants per track and per conference day, float, see Section 4.3.4 |
| Reception | Per cent of sessions with no-show speakers | Percentage of sessions with no-show speakers, see Section 4.3.4 |
| Reception | Number of messages posted | Total number of messages posted by participants in the communication channel, integer, e.g., Slack, see Section 4.3.4 |
| Economic Aspects | Average attendance prices | The mean cost of attending the conference for all participants, calculated over previous years, float, see Section 4.3.5 |
| Economic Aspects | Average internet costs | The mean internet costs for all the participants of the conference, calculated over previous years, float, see Section 4.3.5 |
| Economic Aspects | Average traveling costs | The mean travel costs of all participants of the conference, calculated over previous years, float, see Section 4.3.5 |
| Economic Aspects | Average housing costs | The average housing costs of all participants of the conference, calculated over previous years, float, see Section 4.3.5 |
| Economic Aspects | Average catering costs | The mean costs of catering for the conference, calculated over previous years, float, see Section 4.3.5 |
| Economic Aspects | Average return of funds for research | The mean returns of funds for research for the participants of the conference, calculated over previous years, float, see Section 4.3.5 |
| Economic Aspects | Total price trend | Price trend of the total costs of a conference, calculated on previous years, in percent, see Section 4.3.5 |
| Ecological Impact and Sustainability | Carbon footprint | The total carbon footprint of a conference in kilogram carbon dioxide emissions per day of the conference, see Section 4.3.6 |

| Evaluation Criterion | Metric | Description |
|---|---|---|
| Ecological Impact and Sustainability | Sustainability | The ranking that the sustainability chair or another expert attributes to the conference is based on three levels: highly sustainable, sustainable, and not sustainable, see Section 4.3.6 |
| Ecological Impact and Sustainability | Green innovativeness | Total number of utilized measures to reduce the carbon footprint in comparison to previous years, integer, see Section 4.3.6 |
| Privacy and Security | Threats | The total number of threats identified with a cybersecurity standard, domain-dependent on selected standard, e.g., NIST or ISO 27001, see Section 4.3.7 |
| Privacy and Security | Level of security | The self-assessment score according to the selected standard, domain-dependent on selected standard, see Section 4.3.7 |
| Privacy and Security | Compliance to media guideline | Self-assessment based on best practices with three levels: full compliance, partial compliance, and not compliant, see Section 4.3.7 |
| Privacy and Security | Compliance to GDPR | Self-assessment according to guidelines with three levels: full compliance, partial compliance, not compliant, see Section 4.3.7 |
| Modality | Virtuality | Degree of virtual parts of the conference in per cent, see Section 4.3.8 |

**Table 4.1:** Table of Sample Metrics

## 4.4 Dimension Three: Collection of Data for Metrics

In this section, we delve deeper into data collection and its associated challenges and risks in order to finally answer RQ2 on what metrics can be used for evaluating virtual conferences in combination with RQ3 on how we can collect and compare the necessary data. Additionally, we need to determine how to collect data for our reference framework. First, we will survey elementary definitions and define scales and mapping tables to provide us with the necessary means to refine the sample metrics used for our reference framework. Then we will investigate the techniques used to collect data in order to select the right ones for our approach in the practical implementation. To do this, we need to ensure that the state of the data is appropriate for our purposes, which means cleaning the data and transforming

it if necessary. Only then can further processing take place and meaningful data analysis be performed.

In addition to the basic techniques for collecting data, we will also look at automated and simulated data collection. Special attention will also be paid to the data already collected by tools like Zoom or Slack. Another data collection measure is monitoring the attention and interest of the audience through measurements based on facial recognition. However, this is questioned in terms of privacy violation.

Especially in light of the global climate crisis and the high carbon footprint of conferences, the challenge of collecting data for green metrics is also addressed and further metrics are derived from this research in Section 4.4.10. Another main topic area is the risks of data collection and the issues related to privacy, usability, and accessibility, for which we want to raise awareness for the organization and research of future virtual conferences. Related to this, ways to secure sensitive data will also be highlighted. Additionally, how existing guidelines and standards can be used to protect personal data and comply with the requirements of the GDPR will be explored.

After the groundwork is laid, we will decide which approach we will refine the sample metrics to include our newly gained insights from this section. Finally, we present our proposal for an Evaluation Criteria Fingerprint based on our findings in Section 4.3. With this fingerprint, it is possible to compare virtual conferences and their characteristics visually.

### 4.4.1 Definitions

It is necessary to introduce some definitions to ensure a standard level of knowledge to provide a basis of understanding for our study of data collection methods.

#### Metric

The entry for the term "metric" in the MWOD is coined as *a standard of measurement*[17]. Ellingwood, as we informed in Section 2.2 describes metrics as raw measurements of resource usage or behavior that can be collected and observed throughout the system [Ell17]. According to Kounev et al., [KLK20], a metric corresponds to a value derived from measurements and measures. We expand upon the preliminary definition that we utilized in Section 2.2 to fit this section's more advanced requirements and define "metric" as follows: *A metric is a value that is obtained by applying rules or measurement processes or utilizing measures, which summarizes observed properties of events and objects.* In contrast to a composite metric, we may refer

[17] `https://www.merriam-webster.com/dictionary/metric`, accessed: September 2022

to a single metric as an elementary metric. We define a composite metric as a combination of several elementary metrics or composite metrics to achieve a cognitively more processable quantity.

### Measurement Scales [KLK20]

In order to understand how we need to further process data to obtain metrics and ultimately answer RQ3, we need to look at the available mapping functions and scales for measurements.

| Scale | Mapping | Operations | Statistics | Example |
|---|---|---|---|---|
| Nominal | Unordered | =, != | Mode, frequencies | Conference types: Hybrid, physical, virtual. |
| Ordinal | Adds Ordering | <, > | Median, percentiles | 0 = low attendance, 1 = medium attendance, 2 = high attendance |
| Interval | Adds distance function | +, - | Mean, standard deviation | Temperature, time, dates |
| Ratio | Adds unit and zero point | *, / | Geometric mean, coefficient of variation | 1 h = 3/4 h + 1/4 h |
| Absolute | Adds natural unit | | | 100 sheets of paper, 20 grade A conferences |

**Table 4.2:** Measurement Scales, adapted from Kounev et al. [KLK20]

We distinguish five measurement scales: nominal, ordinal, interval, ratio, and absolute. In Table 4.2, we have listed all five measurement scales with their corresponding properties. Each of the scales is more informative than before, i.e., mathematical operations not applicable to the previous scales are now newly added. The same applies to the mapping functions, e.g., an interval scale includes both order and distance functions to measure the interval between two values.
A *nominal scale* is used to assign qualitative values, e.g., categories or names with no inherent ordering, and cannot be mathematically computed. Measurements that are based on this scale are called *categorical* or *qualitative data*.
An *ordinal scale* allows for comparison between values since it assigns an ordering to its values. However, there is no inherent meaning of the relative position of the values, i.e., ordinal scales do not include a ranking. In the example in Table 4.2, we have defined three values in terms of attendance of participants at a conference. What exactly does low attendance signify,

and how much difference is there to high attendance? We know that high attendance is probably better than low attendance, but not how much better.
A solution for this problem is the *interval scale*, which adds a distance functionality and meaningful intervals to its repertoire. The difference of values on the scale is quantified by scale points so that the relative distance of values can be measured.
The following scale is the *ratio scale*, which adds a meaningful zero point, and the same ratio between values carries everywhere on the scale the same meaning as the interval scale. An example of a ratio scale is temperature, e.g., the Kelvin scale or time. One hour can be divided into 1/4 h plus 3/4 h, which allows for comparing values all over the scale since the ratio signifies the same meaning, no matter the starting point. Many measurements from physical sciences and engineering are based on the ratio scale, e.g., mass, length, volume, time duration, or power consumption.
Lastly, the *absolute scale* is a particular case of a ratio scale based on a natural unit, representing a measure to count discrete objects or probabilities. On an absolute scale, no transformation is allowed other than the identity transformation. For example, 20 sheets of paper have a natural unit, and we can easily up the value to 100 sheets of paper. We can still swap each sheet with another and count them as identical. For another example, we can consider the classification "grade A conference ."We expect from each top conference a high set of standards, and we would expect the same from each conference in the category, regardless of their specific fields of expertise. Additionally, the ranks of conferences and the differences between them are meaningful. They can be measured through the various metrics that contribute to the overall ranking, e.g., the h-index, which can be applied to authors or journals and measures the number of papers published that have been cited at least h times[18].

### Qualitative Rating Scales

As anticipated in Section 2.2, the Likert scale is a rating scale used in the field of qualitative evaluation [MB16]. Two poles, from absolute disagreement to absolute agreement, are represented on a scale of five to ten points [MB16]. In the evaluation, the range on the one hand and the frequency on the other are decisive [MB16]. It is particularly important to offer a moderate or neutral answer option in the middle, as the result could otherwise be distorted if users have to choose one extreme [MB16]. Even though an evaluation of Likert scales is best done by mapping them to numerical values, e.g. from 0 to 4, it is recommended to verbalize the answer options themselves, e.g. "completely agree" [MB16]. Another metric based on a similar rating scale is the Net Promoter Score, which measures the willingness to recommend a product or website using an eleven-point scale from 0 to 10 [19]. Detractors or critics are users

[18] `http://www.conferenceranks.com/`, accessed: September 2022
[19] `https://www.ionos.at/startupguide/produktivitaet/nps-was-ist-der-net-promoter-score/`, accessed: September 2022

who select values in the range 0 to 6. Users who select 7 or 8 as an answer are classified as indifferent. Only users who select response options greater than 8 are classified as promoters.

### 4.4.2 Reasons for Data Collection

When the amount of data collected has long exceeded human comprehension, it is reasonable to ask how more extensive data collection can be justified, especially in light of the climate crisis.
The improvement in an area needs to uncover those problems that need to be solved in the first place. Without sound data, these problems may remain undiscovered or be looked for in the wrong places. Researchers must rely on good data if they are not to act on mere opinion.
Since conferences are an essential part of knowledge sharing for researchers, not only does the research need to be based on sound data, but the conferences themselves need to be analyzed and optimized.
One such attempt was undertaken by Sarabipour et al. In their 2021 study, Sarabipour et al. [Sar+21] examined 270 physical conferences for their environmental impact, inclusivity, and accessibility features. Their analysis indicates that the shift to hybrid or entirely virtual conferences are inevitable due to environmental reasons. It also makes sense for other reasons, such as increased inclusivity and a lower barrier to participation. For example, only 6 % of the 270 conferences claimed to have any form of green policy. Additionally, it was discovered that the total amount of generated carbon dioxide amounted to more than 2 million tons, and 35 % of these could be attributed to conference attendance (including air travel), and 20 % amounted to the infrastructure of the conferences. These are compelling reasons to devote more action to adopting environmental-friendly policies, which are grounded in data, and to consider a switch to virtual conferences.
This argumentation provides an excellent reason to collect data on conferences and enable continuous, automated evaluation of them through suitably selected metrics. In this way, we contribute to transforming conferences, as virtual knowledge exchange platforms, into an environmentally sustainable form that excels in the other evaluation criteria.

### 4.4.3 Basics of Data Collection

[Fox21]

When collecting data, one should rely on high-quality data from the outset and not be under the illusion of being able to achieve the desired quality later through transformations, as this is a tedious and time-consuming task and may lead to errors due to muddled data. The source


**Figure 4.5:** Data Analytics Process [Fox21]

data must already meet several requirements. Therefore, to ensure that data has the required quality, the following criteria must be observed [Fox21]:

- Accuracy: Measurements or metrics must correctly portray what has been observed on events and objects.
- Relevance: The selected data must relate directly to the studied events and objects.
- Representative: Data types must be selected to reflect the studied events and objects as faithfully as possible.
- Well-defined: The meaning of the data must be precisely defined in the form of a schema, metadata, or a data dictionary.
- Complete: The selected data must include all potentially relevant measurements and metrics.
- Granular: The selected data types must be defined over sufficiently large intervals and a sufficiently large level of detail to reflect the whole variance of the data.

Foxwell et al. [Fox21] defined the typical data analysis process and divided it into several granular steps, see Figure 4.5.

#### Design and select data representation

Before a data set is selected or generated, Foxwell et al. already advise determining the desired representation of the data and how to visualize it. In Section 4.4.1, we defined five different measurement scales. Now we map these scales to data types and visualization techniques:

In this section, we address the question of how we should collect data for our reference framework, what kinds of data sets are suitable, and how data sets should be prepared for further processing them in our evaluation tool. Furthermore, this information is also important for future virtual conferences and should contribute to a better evaluation and dissemination of well-prepared data sets. The latter point is especially important because currently there is a lack of data for virtual conferences and therefore the starting point is rather poor to compare them. Foxwell et al. [Fox21] stress that only values should be used in the data set and not their

| Scale | Data Type | Visualization |
|---|---|---|
| Nominal | Character, string | Bar chart, pie chart |
| Ordinal | Character, string | Bar chart |
| Interval | Integer, float, double | Boxplot, histogram |
| Ratio | Integer, float, double | Boxplot, histogram |
| Absolute | Integer, float, double | Boxplot, histogram |

**Table 4.3:** Mapping of scales to data types and visualizations [Fox21]

units of measurement, and, most importantly, no units of measurement should be mixed. Instead, the units of measurement should be defined in a data dictionary. Pie charts should be avoided if possible, as they can be visually misleading and are inferior to bar plots in terms of clarity.
Time formats are a common source of errors and should therefore be handled with extra care. One way to avoid time format conversion errors is to store the duration in HH:MM:SS format. Another type of data prone to conversion errors is data related to currency and money, as spaces or commas can be used instead of a decimal point in various places, depending on the source country. Therefore, Foxwell et al. [Fox21] advice standardizing the quantities at the start of the project, as correcting such conversion errors is a time-consuming and effort-intense process.
Additionally, good quality data is expressed in standard units and should be transformed by normalizing it. An appropriate numerical precision must be selected when dealing with interval and ratio data. Furthermore, issues with multicollinearity should be avoided, i.e., measurements or metrics that closely measure the same qualities and are closely correlated should not be used in the same predictive model. However, multicollinearity is not always apparent before a thorough dataset analysis and can be discovered later. Finally, non-numeric numbers should not be used in measures but instead converted to a numerical data type.
In summary, good data sets should contain an adequate amount of relevant measurements that have been checked not to include any conversion errors or unnecessary units and are defined in fitting data.

#### Generate or obtain dataset

From the outset, it is necessary to have an idea of what data, data collection tools, and methods will be used for analysis to obtain a good data set. Foxwell et al. [Fox21] argue that envisioning the future analysis should accompany the selection of data representations. However, this does not mean that potential biases should be ignored. The critical point is that data and anal-

ysis should be shown because simply listing statistics does not promote understanding of the findings.
Therefore, fundamental analyses should be conducted on the selected data as a first step. Data that can be measured on the nominal or ordinal scale are called qualitative data. Counts and frequencies are needed so that visualization can use these data further. Data that originate from an interval, ratio, or absolute scale must be made interpretable with means, ranges and variances.
After describing the data, relationships between the individual measurements and metrics must be identified or established. It must be questioned whether it is possible to group them and create compound metrics. It is essential to look for patterns already in this step. Furthermore, it is necessary to be aware that correlation does not equal causation because this is a common source of error.
Datasets should be shared with other researchers to contribute to knowledge dissemination. For this to succeed, the data must be documented and structured. The following criteria must be applied to a dataset:

- Findable: The dataset should be uniquely identifiable, completely described, and made publicly available.
- Accessible: The dataset must be able to be retrieved with standard techniques.
- Interoperable: The data representations should be chosen so that they can be compatible with other datasets.
- Reusable: The dataset must be easily reusable in other research endeavors and contain metadata and a guide or rules for using it.

Apart from a sufficient description of the dataset, a dataset dictionary should contain not only a description of the data but the owner or creator and the place of origin. Good metadata contains as much detail and as many descriptions as needed for clarity. The naming of measurements and metrics must be as precise and meaningful as possible. The dataset should also have a name that can be easily found. Standard formats should be chosen for dissemination of the data set. In this case, .csv or .json files prove to be the best.
Finally, you should ask yourself whether the dataset is ready for distribution. There are three main factors to consider in this regard:

- Whether the data set is saved in a standard format.
- Whether the dataset can be loaded into the common analysis software.
- Whether the dataset can appropriately answer the research questions for which it was collected.

If these conditions are met, the dataset can be further processed.

### Clean and transform data items

A good practice is to inspect the data visually, e.g., with box plots and histograms, to assess the quality of the data set. Furthermore, it is advisable to analyze the data based on their definitions, e.g., their possible range, to detect outliers. Three main challenges for data cleaning are the following:

- Human error in the data collection and data post-processing process
- The handling of missing data
- Understanding the sources and meaning of outliers

These challenges can be addressed in the following ways:

- Visually: Data can be visually inspected by viewing it with appropriate tools.
- Programmatically: Coding of error discovery checks, e.g., range or null values checks or utilizing libraries.
- Automated: Utilizing error discovery and correction systems.

If possible, data cleaning should be done by multiple people and as an iterative process. Nevertheless, data cleaning is a complicated and time-consuming process that requires some accuracy and persistence. It may be necessary to define acceptable ranges of values in which values should still be considered valid.

### Explore and summarize data

Based on the guiding idea for the analysis, this step transforms the data using suitable software or libraries. Compound metrics are created, and qualitative data are summarized so they can be interpreted. If not already done for the data cleaning process, the data is now prepared visually. The measurements and metrics are examined based on domain knowledge, and interpretation skills and relationships are detected.

### Further data analytics steps

In the last three steps of Figure 4.5, the meaning of the data is determined and tested along with formulated hypotheses. Based on these hypotheses and their applicability, conclusions can be drawn and turned into research results. However, these steps are beyond the scope of this master thesis, as our evaluation system for virtual conferences only provides the basis for further analysis.

### 4.4.4 Collecting Data from Tools

In this section, we investigate the following questions: How can tools be used to improve and supplement the data collection process for evaluating virtual conferences? Are there any viable tools for evaluating usability so that we can add the collected data to our sample metrics or do we have to rely on user testing? When assembling a technology stack for virtual conferencing, it is an efficient practice to use as few tools as possible while covering as much of the desired functionality. It is also essential to consider which metrics can already be covered by the selected tools. Ideally, the tools and their collected data can already be used for data analysis without any significant further processing. Therefore, when selecting tools for virtual conferences, it is also essential to consider how they can be used to evaluate the conferences.

Typically, to hold a virtual conference, one will need a tool for video and audio transmission, another tool for the paper submission and review process, and at least one tool that enables social interactions among conference participants. Apart from this, a website will also be used to collect web analytics. Depending on the different orientations, other tools may be used, for example, additional tools for social exchange or several streaming platforms. Also, a tool for archiving video recordings taken during the conference could be used. For example, if Zoom is used for video conferencing, the Zoom REST API[20] can be used to export and reuse metrics and reports. In particular, quality of service metrics for individual users can be queried and need not be monitored by extra tools. For social interactions, for example, Slack can be used as a messaging tool that offers persistent channels and groups and private direct messaging. For social interactions, for example, Slack can be used as a messaging tool that offers persistent channels, groups, and private direct messages. While Slack does not have a REST API, it does at least have a Web API with HTTP RPC-style methods that can be used to query JSON reports[21]. It is also possible to create custom Slack apps and further configure monitoring using Bolt or the Java SDK. Data and metrics collected by a paper submission and review tool can also be used for further processing and evaluation. This data can be added later via ingest without API or query capability.

Although it would be obvious to use tools for automated web usability testing, it can be discouraged due to the current state of popular tools for these purposes. Namoun et al. [NAT21], in their review of automated website usability evaluation tools, found that the results of the tools are questionable due to several shortcomings. First of all, they only examine a small number of aspects that constitute usability. They also focus on areas that should not play such a crucial role in usability, such as search engine optimization (SEO) or performance. That being said, there are significant variations in the scores of the same tools for the same websites.

[20] `https://marketplace.zoom.us/docs/api-reference/zoom-api/methods,accessed:September2022`
[21] `https://api.slack.com/web`, accessed: September 2022

Furthermore, interpreting the results is difficult since references and benchmarks for the determined scores are often missing, or the technical jargon is challenging for non-specialists. It is also particularly critical that the specific web usability issues are not named, and many tools do not make any recommendations on how overall usability can be improved. Therefore, we have to rely on user testing to gain valid insights into usability.

In any case, the following recommendations can be made for data collection with tools for evaluation purposes of virtual conferences:

- As few tools as possible should be used that complement each other as well as possible.
- Metrics and data already collected by tools should not be collected twice if possible, except for evaluation purposes of the tool used.
- Evaluation criteria and their associated metrics that are particularly well-supplemented by data from tools include technological background and reception.
- If possible, the reporting mechanics and APIs of the tool should be used directly, and tools where this is possible, should be preferred to those that do not offer this possibility.

### 4.4.5 Simulated Data Collection

It is crucial to test the performance capacities of websites and web services sufficiently in advance so that a virtual conference can occur unhindered and is equipped for the expected number of users. For this reason, it is recommended to perform load tests, e.g. with a tool such as Apache JMeter[22] or Artillery[23]. There are also other reasons for using such tools. Since it is not the performance of a website or web service under static load that is relevant, but rather its behavior under dynamically changing circumstances, it is possible to create scenarios with Artillery or workloads with JMeter to reflect such dynamic behavior.

Simulated data collection consists of generating a data set using the tools mentioned above. Hasnain et al. [Has+21] argue that QoS metrics based on simulated data sets are only marginally different from those collected in real-time and can therefore be utilized for evaluating a virtual conference.

[22] `https://jmeter.apache.org/`, accessed: September 2022
[23] `https://www.artillery.io/`, accessed: September 2022

### 4.4.6 Automated Data Collection

So far, we have assumed that the data is collected and maintained manually. However, we see potential in automation to relieve researchers in terms of time and energy and enable continuous evaluation. In this section, we will discuss if automation is feasible for evaluating virtual conferences and what kind of automation we could use in our reference framework. Neither Hohlfeld [HGD21] nor Vahdati [Vah+21] with their related work considers the possibility of automatically collecting or evaluating the data gathered at the conferences. Admittedly, analyzing data can be challenging, and not every metric collected may be suitable, primarily if free text fields are used, which cannot be easily analyzed, at most, with sentiment analysis. Also, metrics for which no clear standard can be defined and thus their meaning must be interpreted by an expert are only suitable for automated evaluation to a limited extent. Despite these limitations, automation facilitates the work of researchers by allowing standard tasks to be performed without further intervention, leaving only the results to be evaluated. Ideally, a system is set up so that only occasional maintenance is required. Most data collection and cleaning are already done in advance so that researchers can restrict themselves to the analysis process.

Suppose automation is to be used to evaluate virtual conferences. In that case, it is essential to note that the data sources used are likely heterogeneous and need to be brought to a common denominator. In addition, the data must be cleaned before it is integrated into the database, as subsequent cleaning would be more challenging to accomplish. When selecting data sources and sufficiently high quality, availability and how consistent they are is also essential. If data sources change regularly, they can become more error-prone, and new edge cases must be considered for the automated system. The database must be checked from time to time, and the functionality of the automated system. One possibility is to define metrics and test cases for the automated system in addition to manual inspections, thus creating a continuous evaluation.

One option for implementing automation is to use CronJobs, which are jobs executed at a repeating schedule and can be controlled using Kubernetes[24], for example. Another alternative is using scheduled tasks with annotations in Spring Boot, which is also based on cron expressions[25].

### 4.4.7 Measuring Engagement

Engagement is an important principle to describe how much involvement a particular product achieves for the user. In our case, the question is how involved participants of conferences

[24] `https://kubernetes.io/docs/concepts/workloads/controllers/cron-jobs/`, accessed: September 2022
[25] `https://www.baeldung.com/spring-scheduled-tasks`, accessed: September 2022

are in exchange with others and their participation in the talks. Especially for conferences that are supposed to enable scientific exchange, which is more difficult in virtual conferences without further measures, it is necessary to have a high level of user engagement. In addition, high user engagement is a quality feature of conferences, indicating a better reception.

In their systematic review of 351 articles and 102 definitions, Doherty et al. [DD18] captured the state of research on engagement. Based on their research, we define engagement as an immersive process that reflects user involvement and interaction. According to Doherty et al., [DD18], subjectivity-oriented approaches on the one hand and objectivity-oriented approaches on the other are ideally combined to measure engagement. Among the subjectivity-oriented approaches are the following:

- Questionnaire methods, e.g., employing a user engagement scale
- Experience sampling methods, e.g., repeated experience reports at the moment
- Observational methods, e.g., ethnographic or qualitative analysis of engagement
- Inferring engagement from subjective data, e.g., assessing the bias of collected subjective data

Among the objectivity-oriented approaches can the following be counted:

- Behavioral trace methods, e.g., measures such as the number of mouse clicks on a web page or dwell time on websites
- Psycho-physiological measures, e.g., electrocardiography
- Inferring engagement from objective data, e.g., expert ratings of collected objective data

Conference organizers should combine measures from the two approaches to have a clearer picture of user engagement. For our practical project for the master thesis, it is suitable to combine, for example, questionnaire methods and collected activity metrics by tools.

Another aspect we want to mention is the AI-based emotion recognition considered by Zoom[26]. Although it may be tempting to be able to collect data in this way and determine which content evokes which emotion in users and when their attention drops, we see this as a significant risk to user privacy. Even if the data is anonymized, there is still a risk, especially if the number of users is small, that attribution can occur and that video data can be reused for other purposes not disclosed by Zoom. Such continuous analysis of emotions requires records on which AI-based calculations can be performed.

[26] `https://www.biometricupdate.com/202204/reported-ai-based-emotion-recognition-by-zoom-irks-rights-advocates`, accessed: September 2022

Another example is De Carolis et al. [De +19], who analyzed student engagement in their study using behavioral cues from facial expressions, head movements, and gaze behavior. In order to assess learning effects and engagement, this method of analyzing engaged faces would be applied. Again, we argue that while this data may help assess engagement and the quality of instructional materials outside of a test series, it violates users' privacy rights. Furthermore, according to Doherty et al. [DD18], there is a risk that such behavioral cues will be misinterpreted, as they may be emitted differently depending on ethnicity and also one's personality.

### 4.4.8 Challenge of Collecting Data for Green Metrics

Given the urgent need to collect more data in the area of ecological impact and sustainability in virtual conferences to ensure the effectiveness of measures and comparability among conferences, we asked ourselves what approaches we could use to address this issue. Are there other green metrics that we can include and how can they be collected? These questions concern us not only for our reference framework, but also beyond that for the future work of evaluating conferences.

Although virtual conferences have lower carbon dioxide emissions than their physical counterparts, there is still a need for optimization to reduce the carbon footprint [Wil+21; Sar+21; Pie+20]. As Faber et al. [Fab21] noted in their literature review, most researchers still focus on virtual meetings rather than virtual conferences, which have many more participants and a more complex ecosystem. However, it is challenging to determine a virtual conference's carbon footprint.

It is difficult to determine which factors must be considered since indirect factors contribute to a higher carbon footprint. An example of a factor that cannot be calculated with certainty that has a profound effect on carbon emissions is the embodied energy of Internet network infrastructure [Fab21]. An additional problem is the collection of green metrics. Especially for virtual conferences, eco-friendly measures should already be part of the policy, as data collection is complex and should be considered from the beginning, for example, if an LCA is to be performed [Wil+21]. Faber et al. [Fab21] identified the following relevant factors for their evaluation framework of carbon emissions at conferences:

- Energy use from network infrastructure
- Computer energy use
- Computer embodied energy, i.e., the sum of all energy used to produce a computer
- Actions performed when videoconferencing

These metrics are a start but do not capture the emissions of virtual conferences in all their complexity, e.g., physical and organizational meetings, including flights and lodging. However, they cover significant contributing factors to the carbon footprint of virtual conferences.

Since calculating the carbon footprint of a virtual conference is a highly complex task, it is advised to use automatization for as many sub-areas as possible [Wil+21]. Another complication is the lack of research regarding virtual conferences [Wil+21; Bon+22], their sustainability, and how exactly the carbon footprint should be determined. There are no comparisons yet of which virtual conference formats have proven to be more environmentally sustainable than others [Wil+21].

### 4.4.9 Risks of Data Collection

In this section, we focus on the risks associated with data collection to mitigate these risks in the implementation of our reference framework and to provide guidance for organizers of future conferences. These risks are manifold and cover different areas, such as security breaches according to GDPR article 32, disclosing sensitive data, unlawful processing, using cloud services without complying with privacy principles, and endangering the rights of data subjects. Furthermore, we deal with the security of sensitive data. In addition, we are addressing the protection of data and GDPR compliance with the help of security standards, e.g., NIST or ISO 27000.

#### GDPR

The GDPR is an EU data protection and security law enforced since May 25, 2018. It applies not only to EU citizens and residents but also when data of EU citizens or residents are processed, even if the location of the data processor is not in the EU[27]. In addition, severe penalties are imposed for non-compliance with the law. In particular, the GDPR strengthens the rights of data subjects, which include the following[28]:

- The right to be informed
- The right to access
- The right to rectification
- The right to erasure
- The right to restrict processing

[27] `https://gdpr.eu/what-is-gdpr/`, accessed: September 2022
[28] `https://gdpr.eu/what-is-gdpr/`, accessed: September 2022

- The right to data portability
- The right to object
- Rights concerning automated decision making and profiling

Organizers of virtual conferences should be considerate of the GDPR, especially when using an evaluation system that is supplied with data sets containing personal or even sensitive data. It is first necessary to distinguish between personal and sensitive data to determine which data needs special protection and how compliance with the GDPR guidelines is possible. According to Article 4 of the GDPR, personal data includes the following[29]:

- Name
- Identification number
- Location data
- Online identifier
- Factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of a natural person

Personal data can then be defined as any information that relates to a natural person that can be identified directly or indirectly by this information. Sensitive data, on the other hand, is a particular category of personal data, which includes the following areas according to article 9 off the GDPR[30]:

- Ethnic or racial origin
- Political opinions
- Religious or philosophical beliefs
- Trade union membership
- Genetics
- Biometrics
- Health
- Sexual activities and sexual orientation

[29] `https://gdpr-info.eu/art-4-gdpr/`, accessed: September 2022
[30] `https://gdpr-info.eu/art-9-gdpr/`, accessed: September 2022

A risk can already arise if the consent of the participants of a virtual conference has not been explicitly obtained for the data collection. Those who manage data, the data controllers, are fully responsible for the personal data and must be able to prove that they comply with the GDPR[31]. In addition, care must be taken to ensure that data subjects not only have a right to be forgotten but also to access their personal data and request changes if necessary. Therefore, it is recommended that a Data Protection Officer (DPO) be appointed to take responsibility for and oversee the proper handling of personal data. Data subjects can then contact the DPO directly and request the implementation of their rights [CMR21].

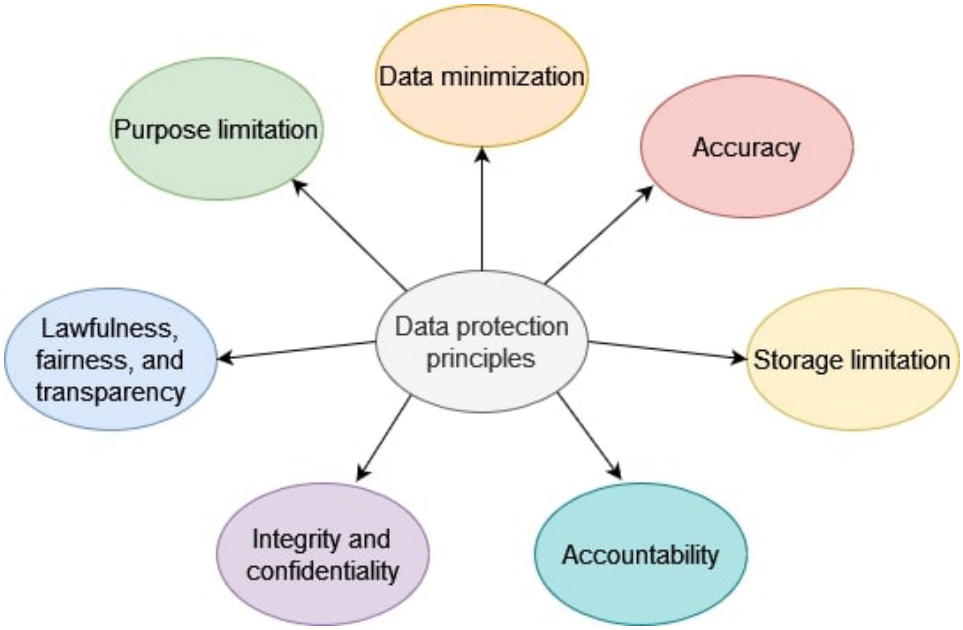



**Figure 4.6:** Data protection principles according to GDPR

In order to bring about a guarantee of proper processing of personal data, it is necessary to adhere to the data protection principles defined in Article 5 of the GDPR[32], which can be seen in the graphical overview of Figure 4.6. It must be ensured that the data is processed in a lawful, fair, and transparent manner. The data may also only be used for explicit and legitimate purposes, meaning the purposes must be limited. Personal data should also be adequate, relevant, and tailored to the purposes of the processing, which implies data minimization. In addition, this data must be kept accurate and, when applicable to the purposes, up to date. Any data not meeting this standard must be deleted or rectified without delay. In terms of limiting the data storage, it is imperative that the data is only kept in a form that allows personal identification for as long as necessary for the processing. The data controller is responsible for compliance with the GDPR and must be able to prove that it complies with it at all times, which is in line with the principle of accountability. Finally, the data must be processed in a manner that guarantees the security of the personal data and provides

[31] `https://gdpr.eu/what-is-gdpr/`, accessed: September 2022
[32] `https://gdpr.eu/article-5-how-to-process-personal-data/`, accessed: September 2022

protection against unauthorized access, improper processing, accidental loss, damage, or destruction. Appropriate technical or organizational measures (TOM) should be used for this purpose.

According to Utzerath et al. [UD21], by the end of June 2021, one-fifth of all GDPR breaches have occurred due to Article 32 violations. According to Utzerath et al., by the end of June 2021, one-fifth of all breaches of GDPR occurred due to Article 32 violations. No sufficient TOM had been deployed in all cases that showed such violations of Article 32. Since the lack of sufficient TOM alone makes for a violation, and any organization can experience a data or cyber breach at any time, it is imperative to focus on deploying proper TOM.

#### Privacy, Accessibility, and Usability Issues

Beyond the GDPR, there are other issues related to data collection that we want to raise awareness for organizers of future virtual conferences and researchers that evaluate the data on these conferences. Alwabel et al. [Alw20] make an additional distinction in the collected data types, as they see a risk in terms of the identifiability of an individual through the information that does not fall under personal data. Non-personally identifiable information is any information that can be seen as private and related to an individual's life. Three main areas are affected:

1. Social life: Information that includes not only sensitive data as defined by the GDPR but also financial data and job performance information, as well as hobbies and interests related to entertainment and commercial products.

2. Cognitive/Expert life: Any information that can be used to identify an individual's expertise, such as knowledge, professional background, skills, and professional interests, as well as individual goals and intentions, may be considered sensitive data.

3. Digital life: Any information that can be used to determine an individual's use of digital data, especially the activities performed on a unique device. Repeated behavior of an individual can also be used to create profiles, e.g., concerning viewing the history or use of a product. It also includes unique, trackable data such as IP addresses or Radio Frequency Identity (RFID) tags.

We, therefore, advise that this data is also treated as if it were GDPR-sensitive data and that appropriate measures are taken to protect it.

GDPR compliance is also a particular issue when cloud architectures are used, such as IAAS, PAAS, and SAAS. Cloud providers must use mechanisms for network protection, encryption, and notifications of authorities and data subjects [GML20].

A particularly vulnerable group to privacy breaches are individuals with disabilities, who often face barriers in managing their privacy. Tools designed to help maintain user privacy are often inaccessible to individuals with disabilities. Therefore, design guidelines need to be created to help create better privacy tools. According to Wang et al. [WP22], the main challenges in designing accessible privacy are:

- Understanding the users' privacy challenges and needs that are not currently served appropriately.
- Taking into account the considerable variation of challenges and needs of the users.
- Considering the complex and frequently intersectional nature of people's disabilities.
- Realizing that assistive technologies may pose privacy risks under certain circumstances, e.g., disclosing private information by text-to-speech measures.

Some measures that can remedy these risks are accessible authentication, accessible CAPTCHAs, privacy-enhancing assistive features, and available assistive tools that aid people with disabilities in gaining access to content and navigating privacy settings [WP22]. A key direction that accessible privacy has to take, according to Wang et al. [WP22], is the turn towards personalized privacy, which is adapted to the needs and specific challenges that a particular user faces.

For people with cognitive disabilities, privacy and security measures can often pose challenges, e.g., selecting the appropriate answers for cookies and privacy settings or understanding the intricacies of privacy agreements. However, many of the usability factors crucial to an effective, efficient, and effortless user experience are also relevant to all users, not only to those with cognitive disabilities, which is all the more reason to improve on these [Käv+22]. Kavrestad et al. [Käv+22] identified four main themes in which privacy and security can be improved in terms of usability for people with cognitive disabilities:

1. Media diversity: It can be crucial for users, e.g., with challenges such as dyslexia, to have multiple ways in which the information can be accessed. Text-to-speech functionality was especially beneficial in understanding complicated or long stretches of text.
2. Clarity and simplicity: Not only people with cognitive disabilities gained when the design was based on these principles. Processing information can be energy-consuming, and badly worded information can be challenging to process. Therefore it can be recommended to provide information in an easy-to-digest manner, as well as summarize the most important information.

3. Limit cognitive load: Learning and re-learning are energy-intensive processes that should be limited to the areas in which it is necessary to learn and memorize. Regarding tools, as little time to learn how to use them as necessary is the directive, and the interaction should also be kept to a minimum.
4. Discriminating design: Some tools are simply designed so that they are impossible to use for specific user groups. Examples of such design faults are hidden input text fields, captchas that require good vision and cognitive processing capabilities, and text elements with small font sizes.

#### Securing Sensitive Data

There are several ways in which sensitive data can be secured, and it is highly dependent on the use case and which measures should be adopted. One crucial factor, in general, is ensuring privacy and security through appropriate measures. This factor concerns storing sensitive data and all steps of processing and transferring the data per the data protection principles. Privacy should be included as a technical requirement to achieve this when developing new software or setting up new IT systems [Dia+22]. Data Protection Impact Assessments (DPIA) should consist of a systematic description of processing activities and their purposes, an assessment of the necessity and adequacy of these activities, a risk assessment of the rights of data subjects, and projected risk mitigation measures. Some TOM that can be recommended include [Dia+22]:

- Pseudo-anonymization and encryption
- Ensuring privacy, integrity, availability, and reliability
- Access to and restoration of data in case of an incident
- Testing and evaluation of protection measures

Additionally, privacy-enhancing technologies (PETs) should be employed to mitigate the risk of data and security breaches since unauthorized data access can lead to severe privacy violations [Sea22]. The next PETs can be recommended [Sea22]:

- Using HTTPS on all websites
- Not mixing HTTP and HTTPS traffic on a single website
- Switching to secure messages that employ the Signal Protocol
- Switching to secure e-mail providers that support TLS
- Deleting old e-mails with sensitive data

- When sensitive data is communicated, authenticating anew is necessary
- Using Two-Factor Authentication
- Using end-to-end-encryption
- Using anonymization, wherever it is feasible

Another option for securing data is to ensure the security of all components involved in the processing and storing of the data with metrics. Viegas et al. [VK22] provide an overview of security metrics for different areas, e.g., endpoint security, network security, or vulnerability management, as well as a guiding principle in the form of a SMART approach when defining metrics:

- Specific: Target the area being measured directly, not a result.
- Measurable: Ensure that the data is accurate and complete.
- Actionable: The data must be easy to understand so that it can be processed without complications.
- Relevant: Focus on measuring what is important in the data.
- Timely: The data must be available when it is needed.

Makri et al. [MGL20] proposed a privacy assessment method for datasets based on data protection principles of the GDPR, for which they offer an array of metrics that require assessments and auditing.

Lastly, sensitive data can be best secured when adhering to relevant standards, guides, or controls. Two standards that can be recommended for this particular use case are ISO/IEC 27001 and ISO/IEC 27002, as well as NIST. However, these standards provide only a good base for privacy and security and should be adapted to meet the strict GDPR. Diamantopolou et al. [DTK20] offers recommendations for achieving GDPR compliance when relying on ISO 27k standards, while Cantiello et al. guide for adapting NIST control to achieve the same means [CMR21].

### 4.4.10 Refinement of Sample Metrics

In RQ2, we formulated the question of what metrics can be used to evaluate virtual conferences. In Section 4.3 we have already created a table with sample metrics, see Table 4.1. However, these metrics still need to be prepared to make them suitable for data collection. We need to determine the form of data that will be collected for the metrics and how can be visualized. This is not only relevant for RQ2, but in particular for RQ3, how the data should be collected and compared. Therefore, in this section, we proceed with the refinement of the

sample metrics based on the findings from the previous sections. Our reference framework allows the addition of arbitrary numerical metrics, making it possible to tailor the evaluation to a specific conference. Additionally, these custom metrics can be visualized with custom tiles and custom charts. For this reason, the refined sample metrics, see Table 4.4, include information on how best to define and visualize the data, based on best practices, see also Section 4.4.1 and 4.4.3. We have mapped the metrics to appropriate scales, data types, and visual representations. Although it is possible to visualize data from a nominal scale as a pie chart, we chose not to do so because, according to Foxwell et al., as discussed in Section 4.4.3, they are visually misleading and weaker in terms of clarity than bar plots. When Likert scales are used, special care must be taken to ensure that, on the one hand, the choices are labeled for the user and, on the other hand, that they are numerically mapped internally so that they can be interpreted in the best possible way, as detailed in Section 4.4.1. As we discussed in Section 4.4.8, more data on virtual conferences in the field of ecological impact and sustainability are needed to assess the effectivity of green measures as well as make virtual conferences more comparable. For this reason, we decided to add four new sample metrics to our refined sample metrics table, see Table 4.4, based on the research of Faber et al. [Fab21]. They reflect the main contributors to a high CO2-footprint for virtual conferences. Another addition is the metric user engagement, which was determined in Section 4.4.7 to be an important factor for measuring the reception of a conference. However, we are conflicted on the potential ramifications of user engagement assessment through tools in terms of privacy, especially in regards to compliance with the GDPR 4.4.9. We advise to use subjective user engagement assessment methods in combination with objective methods that respect the privacy of users, e.g., by utilizing anonymization and other measures to secure sensitive data, as detailed in Section 4.4.9 and 4.4.9. Furthermore, we advise to especially pay attention to behavior patterns that could identify individuals, when the sample size is small, see Section 4.4.9. Although we would have liked to add more usability scores collected with tools to our refined sample metrics, we decided against it, since most usability tools do not provide valid and actionable insights, as detailed in Section 4.4.4. In terms of metrics for the evaluation criterion technical background, we have added two new metrics, concerning the response time and the number of requests per second for the conference website, collected with load testing tools, as detailed in Section 4.4.5. We argue that with the addition of these two metrics, virtual conference websites can be better evaluated for their stability with high loads, as would occur during the holding of a conference.

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Accessibility and Usability | Perceivable score | Based on WCAG 2.1 AA principles, twenty sub-criteria, each fulfilled sub-criterion amounts to 1 point, score consists of the total number of reached points, see Section 4.3.2 | Interval | Float | Boxplot or histogram |
| Accessibility and Usability | Operable score | Based on WCAG 2.1 AA principles, seventeen sub-criteria, each fulfilled sub-criterion amounts to 1 point, score consists of the total number of reached points, see Section 4.3.2 | Interval | Float | Boxplot or histogram |
| Accessibility and Usability | Understandable score | Based on WCAG 2.1 AA principles, ten sub-criteria, each fulfilled sub-criterion amounts to 1 point, score consists of the total number of reached points, see Section 4.3.2 | Interval | Float | Boxplot or histogram |

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Accessibility and Usability | Robust score | Based on WCAG 2.1 AA principles, three sub-criteria, each fulfilled sub-criterion amounts to 1 point, score consists of the total number of reached points, see Section 4.3.2 | Ratio | Interval | Boxplot or histogram |
| Accessibility and Usability | Total Accessibility Score | Based on the four scores for the WCAG 2.1 AA principles, the sum of all four scores, the maximum number of reachable points is 50, see Section 4.3.2 | Interval | Float | Boxplot or histogram |
| Accessibility and Usability | Perceived usefulness | Rated in usability tests, in percent or measured with five-point Likert scale, see Section 4.3.2 | Nominal or ordinal, Likert | Float | Bar chart |
| Accessibility and Usability | Ease of use | Rated in usability tests, in percent or measured with five-point Likert scale, see Section 4.3.2 | Nominal or ordinal, Likert | Float | Bar chart |

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Accessibility and Usability | Task relevance | Rated in usability tests, in percent or measured with a five-point Likert scale, see Section 4.3.2 | Nominal or ordinal, Likert | Float | Bar chart |
| Accessibility and Usability | Enjoyment of use | Rated in usability tests, in percent or measured with five-point Likert scale, see Section 4.3.2 | Nominal or ordinal, Likert | Float | Bar chart |
| Technical Background | Amount of tools | The total number of tools used for the communication channels, the platform, the review, and submission system, and the website of the conference, see Section 4.3.3 | Interval | Integer | Boxplot or histogram |
| Technical Background | Integration | The level of ease of integrating optional tools into the conferencing setup, e.g., connection to social media, select from levels: not flexible, flexible, highly flexible, see Section 4.3.3 | Ordinal | Float | Bar chart |
| Technical Background | Average response time | Mean response time in ms, see Section 4.3.3 | Ratio | Float | Boxplot or histogram |

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Technical Background | Subjective audio quality | Rated audio quality by testers, select level from: not feasible, bad, adequate, good, perfect, see Section 4.3.3 | Ordinal | Float | Bar chart |
| Technical Background | Jitter | The fluctuation of the delay of received packets in ms, collected by Zoom[33] | Ratio | Float | Boxplot or histogram |
| Technical Background | Latency | Amount of packet delay in ms, collected by Zoom[34] | Ratio | Float | Boxplot or histogram |
| Technical Background | Statistics of the common visual platform | Statistics report of the common visual platform, e.g., Zoom, see Section 4.3.3 | Interval or Ratio | Float | Boxplot or histogram |
| Technical Background | Process pipelines | Evaluated efficiency of submission and review process pipelines in per cent, see Section 4.3.3 | Ordinal | Float | Bar chart |
| Technical Background | Response time | Response time of the conference website, median, collected with a load testing tool, e.g., Artillery [35] or JMeter [36] | Ordinal | Float | Bar chart |

[33] `https://support.zoom.us/hc/en-us/articles/204654719-Dashboard`, accessed: September 2022
[34] `https://support.zoom.us/hc/en-us/articles/204654719-Dashboard`, accessed: September 2022
[35] `https://www.artillery.io/`, accessed: September 2022
[36] `https://jmeter.apache.org/`, accessed: September 2022

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Technical Background | Requests per second | Requests per second of the conference website, median, collected with a load testing tool, e.g., Artillery [37] or JMeter [38] | Ordinal | Float | Bar chart |
| Reception | Total number of speakers | The total amount of speakers for a conference, see Section 4.3.4 | Interval | Integer | Boxplot or histogram |
| Reception | Total number of talks | The total number of talks for the duration of the conference, integer, see Section 4.3.4 | Interval | Integer | Boxplot or histogram |
| Reception | Total number of participants | The total number of participants for a conference, integer, see Section 4.3.4 | Interval | Integer | Boxplot or histogram |
| Reception | Average number of questions per talk | The mean number of questions asked via chat for all talks of a conference, see Section 4.3.4 | Interval | Float | Boxplot or histogram |
| Reception | Average number of participants per track per day | Mean number of participants per track and per conference day, see Section 4.3.4 | Interval | Float | Boxplot or histogram |

[37] https://www.artillery.io/, accessed: September 2022
[38] https://jmeter.apache.org/, accessed: September 2022

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Reception | Maximum number of participants per track per day | Maximum number of participants per track and per conference day, see Section 4.3.4 | Interval | Float | Boxplot or histogram |
| Reception | Minimum number of participants per track per day | Minimum number of participants per track and per conference day, see Section 4.3.4 | Interval | Float | Boxplot or histogram |
| Reception | Per cent of sessions with no-show speakers | Percentage of sessions with no-show speakers, see Section 4.3.4 | Ordinal | Float | Bar chart |
| Reception | Number of messages posted | Total number of messages posted by participants in the communication channel, e.g., Slack, see Section 4.3.4 | Interval | Integer | Boxplot or histogram |
| Reception | User engagement | Employing a user engagement questionnaire, based on 5-point Likert scale, see Section 4.4.7 | Nominal or ordinal, Likert | Float | Bar chart |
| Economic Aspects | Average attendance prices | The mean cost of attending the conference for all participants, calculated over previous years, see Section 4.3.5 | Interval | Float | Boxplot or histogram |

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Economic Aspects | Average internet costs | The mean internet costs for all the participants of the conference, calculated over previous years, see Section 4.3.5 | Interval | Float | Boxplot or histogram |
| Economic Aspects | Average traveling costs | The mean travel costs of all participants of the conference, calculated over previous years, see Section 4.3.5 | Interval | Float | Boxplot or histogram |
| Economic Aspects | Average housing costs | The average housing costs of all participants of the conference, calculated over previous years, see Section 4.3.5 | Interval | Float | Boxplot or histogram |
| Economic Aspects | Average catering costs | The mean costs of catering for the conference, calculated over previous years, see Section 4.3.5 | Interval | Float | Boxplot or histogram |
| Economic Aspects | Average return of funds for research | The mean returns of funds for research for the participants of the conference, calculated over previous years, see Section 4.3.5 | Interval | Float | Boxplot or histogram |

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Economic Aspects | Total price trend | Price trend of the total costs of a conference, calculated on previous years, in percent, see Section 4.3.5 | Interval | Float | Boxplot or histogram |
| Ecological Impact and Sustainability | Carbon footprint | The total carbon footprint of a conference in kilogram carbon dioxide emissions per day of the conference, see Section 4.3.6 | Ratio | Float | Boxplot or histogram |
| Ecological Impact and Sustainability | Sustainability | The ranking that the sustainability chair or another expert attributes to the conference is based on three levels: highly sustainable, sustainable, and not sustainable, see Section 4.3.6 | Ordinal | String | Bar chart |
| Ecological Impact and Sustainability | Green innovativeness | Total number of utilized measures to reduce the carbon footprint in comparison to previous years, see Section 4.3.6 | Interval | Integer | Boxplot or histogram |

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Ecological Impact and Sustainability | Network energy use | Energy use from network infrastructure in kilowatt-hours (kWh), as defined by Faber et al. [Fab21] | Ratio | Float | Boxplot or histogram |
| Ecological Impact and Sustainability | Computer energy use | Energy use from computer infrastructure of all participants of a virtual conference in kg CO2-equivalent/computer, as defined by Faber et al. [Fab21] | Ratio | Float | Boxplot or histogram |
| Ecological Impact and Sustainability | Actions performed when videoconferencing | Co2-generating actions, e.g., streaming and accessing digital documents, in kg CO2-equivalent, as defined by Faber et al. [Fab21] | Ratio | Float | Boxplot or histogram |

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Ecological Impact and Sustainability | Computer embodied energy | The sum of all energy used to produce the computer infrastructure that is used by participants to access the virtual conference, in kilowatt-hours (kWh), as defined by Faber et al. [Fab21]. This information is often provided by the manufacturers and can be used to estimate this value | Ratio | Float | Boxplot or histogram |
| Privacy and Security | Threats | The total number of threats identified with a cybersecurity standard, domain-dependent on selected standard, e.g., NIST or ISO 27001, see Section 4.3.7 | Interval | Integer | Boxplot or histogram |
| Privacy and Security | Level of security | The self-assessment score according to the selected standard, domain-dependent on selected standard, see Section 4.3.7 | Ordinal or interval | String or float | Bar chart, boxplot or histogram |

| Evaluation Criterion | Metric | Description | Scale | Data Type | Visualization |
|---|---|---|---|---|---|
| Privacy and Security | Compliance to media guideline | Self-assessment based on best practices with three levels: full compliance, partial compliance, and not compliant, see Section 4.3.7 | Ordinal | String | Bar chart |
| Privacy and Security | Compliance to GDPR | Self-assessment according to guidelines with three levels: full compliance, partial compliance, not compliant, see Section 4.3.7 | Ordinal | String | Bar chart |
| Modality | Virtuality | Degree of virtual parts of the conference in per cent, see Section 4.3.8 | Ordinal | Float | Bar chart |

**Table 4.4:** Table of Sample Metrics

### 4.4.11 Comparison of Conferences: Fingerprint

In this section, we assume that the organizers of a virtual conference selected the appropriate metrics for the evaluation criteria and collected appropriate data for them while holding the conference. How can this data now be evaluated or conferences compared with each other? On the one hand, the data can now be read into the reference framework. With the help of Kibana dashboards, the data can be visually illustrated and interpreted by an expert. With such a dashboard, it is possible to view values for metrics for different conference editions over the years. We will discuss this option in more detail in the practical part, see Section 5 and specifically Section 5.4.4, where we also show a sample dashboard, see Figure 5.6. On the other hand, we would like to offer organizers a way to capture the strengths of their virtual conference at a glance. We believe that a visual solution is ideal to grasp this information

quickly. But how can we show transparently which strengths an arbitrary virtual conference in the field of computer science has to offer? How can we achieve this with our defined evaluation criteria and sample metrics? For the solution of this problem we were inspired by U-Multirank [39]. U-Multirank is a European ranking system for higher education institutions. For the ranking, five dimensions are used: teaching and learning, research, knowledge transfer, international orientation and regional engagement. Furthermore, U-Multirank includes five performance ranks, ranging from A (very good) to E (weak). Instead of providing a composite score, they show the strengths of universities and other institutions and leave it to the individual to select the best one based on their preferences. As our solution should be suitable for various, highly diverse virtual conferences in the field of computer science, we argue that this approach of U-Multirank would be a good fit, especially since it is up to the organizers how important they deem each evaluation criterion and where they would like to improve. Not every virtual conference needs to excel in the evaluation criterion of technical background, for example. However, we think it is necessary that the organizers have this freedom of choice to choose where they want to focus their efforts. To solve the problem, we would like to propose the use of an Evaluation Criteria Fingerprint. The wording fingerprint is meant to reflect that the individual characteristics of a virtual conference are captured. How can we design this Evaluation Criteria Fingerprint and what do we need to define in advance? To reflect the strengths of a virtual conference, we will use the evaluation criteria from Section 4.3. The evaluation criteria already capture essential areas of a virtual conference and show where the focus should be placed during an evaluation. We now assume that experts have selected 3 to 7 metrics for each evaluation criterion. If experts have collected a high amount of metrics, we would suggest to select 3 to 7 core metrics from among them, as we believe this is a reasonable amount to display the strengths of a conference. These metrics could also be taken from the table with the refined sample metrics, see Table 4.4. To ensure that the data are uniformly normalized, we will use a scale of 0 to 1 for performance. We will divide this scale into five performance ranks:

- A: very good, 0.8-1.0
- B: good, 0.6-0.79
- C: adequate, 0.4-0.59
- D: weak, 0.2-0.39
- E: very weak, below 0.2

These performance ranks allow us to highlight areas where improvement is needed, but also to show which areas are already acceptable. We have also mapped them to ranges on the scale. The maximum total score of metrics in a single evaluation criterion is normalized to

[39] `https://www.umultirank.org/about/methodology/our-approach/`, accessed: September 2022

1. We define that the indicator scores of the core metrics that have been selected for each evaluation criterion contribute equally to this total score: $\sum_{n=1}^{i} n/i$, where i is the total number of metrics. For example, if four metrics are chosen for a single evaluation criterion, then each indicator score of a metric contributes 25% of the total score. Each score must be normalized to the scale from 0 to 1. How can the indicator score of a metric be calculated? As arbitrary metrics can be used, an expert needs to define the goal value for a metric and choose how well this goal was reached, e.g., by calculating the distance to the goal value. To summarize, the following steps have to be taken in order to collect the necessary data for the fingerprint:

- Select core metrics for each evaluation criterion
- Assign an indicator score to each metric
- Normalize the indicator scores, so that they are amounting to a total of 1 per evaluation criterion

How can the data be visualized for our Evaluation Criterion Fingerprint? According to Foxwell et al. [Fox21] and our results in Section 4.4.1, a bar plot would be a suitable visual representation for data on an interval scale. Therefore, we have chosen to create a bar plot in RStudio [40]. We have created a sample fingerprint, see Figure 4.7 with generic metrics in the first seven evaluation criteria, excluding the criterion modality, since we are focusing solely on virtual conferences in this master thesis. The criterion modality could easily be added to the fingerprint by assigning it the degree of virtuality as single metric, so that the value between 0 and 1 would directly reflect the amount of virtuality in the conference. In Figure 4.7 we can see on the x-axis the seven evaluation criteria, abbreviated in the following manner: Accessibility (A), Usability (U), Technical Background (TB), Reception (R), EA (Economic Aspects), Ecological Impact and Sustainability (ES), and Privacy and Security (PS). On the y-axis, the level of performance is depicted on a scale from 0 to 1. As the legend on the right shows, the color-coded segments of the bars correspond to the core metrics that have been selected for each evaluation criterion. To evaluate the performance in each evaluation criterion, the performance ranks that we have defined for each range bin on the scale can be utilized. For example, this sample fingerprint shows that the virtual conference achieved a value between 0.4 and 0.59 in the evaluation criterion accessibility, therefore we can assign the performance rank C: adequate. For example, if such a fingerprint would be created for each edition of a virtual conference, it could be seen at a glance in which areas the conference has improved by comparing them visually. The position of the bars suggests how good the performance in the area of a specific evaluation criterion was, while the color coding shows how many metrics were used and how they were rated by an expert. In our sample fingerprint, we can see that the virtual conference was well-rounded, with specific strengths in usability, privacy and

[40] `https://www.rstudio.com/`, accessed: September 2022

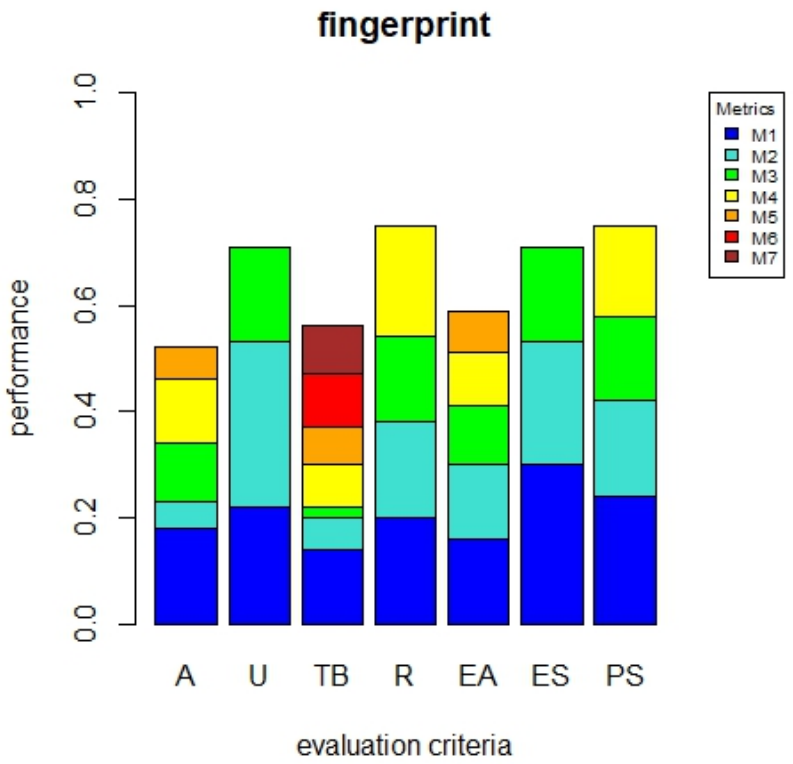



**Figure 4.7:** Sample Evaluation Criterion Fingerprint

security, and it was also well-received, as indicated by the high rank of reception. However, it can be clearly seen by the distance to the top score that there is room for improvement in all areas. As for how viable this Evaluation Criterion Fingerprint is for evaluating virtual conferences, we argue that it allows for an easy comparison of conferences, but that there is still room for improvement, especially since it currently depends on the ratings of an expert. Additionally, automation would improve its usefulness. In any case, the fingerprint can be used to get an overview of the performance of a virtual conference and to interpret the results visually in a simple way. Furthermore, the fingerprint can be used to document the performance of several editions of a conference and to get an overview of how the performance in previous years compared to the current performance. In addition to the dashboarding alternative in the reference framework, the Evaluation Criteria Fingerprint is our contribution on how to better evaluate virtual conferences.

# 5 Practical Case Study: Reference Framework

In this section, the practical part of this thesis is described. To create a user-centered software for evaluating virtual conferences, some prerequisites must be met. Therefore, this section handles the requirements for the framework and what technologies can be used for realizing an appropriate software engineering-based solution. This work should result in a prototype or foundation for a larger framework, but should not be understood as a completed product.

## 5.1 Comparison of Related Work

We first have to analyze tools currently used to evaluate virtual conferences to position our theoretical framework amongst the existing competitors in that area. After an extensive web search in this area, we concluded that no software has full compliance with the needs of our evaluation framework, described before in Section 4.
We then reconsidered broadening the scope so that at least partial compliance with our framework can be analyzed. We identified Bluejeans by Verizon [Ver22] as a solution that also offers evaluation functionality. Two other competitors, a commercial one and a free alternative, were chosen to give an overview of the state-of-the-art in evaluating virtual meetings. Furthermore, we discovered an innovation from research in the form of ZoomSense by Bartindale et al. [Bar+21]. Therefore, in the hope of finding a tool that offers monitoring capabilities, we consulted the Gartner Magic Quadrant for video conferencing tools[1].

### 5.1.1 Bluejeans

First of all, Bluejeans by Verizon [Ver22]. In contrast to Flowmon [Cor20], Bluejeans is an all-in-one package consisting of a solution for hosting video conferencing and monitoring. According to their website, they are mainly targeting businesses with functionality tailored for meeting rooms and integrations to other business-oriented software, like Slack[2]. More interesting in the context of this thesis, a command center is also included. This tool presents

[1] `https://www.gartner.com/reviews/market/meeting-solutions` (accessed: September 2022)
[2] `https://slack.com` (accessed: September 2022)

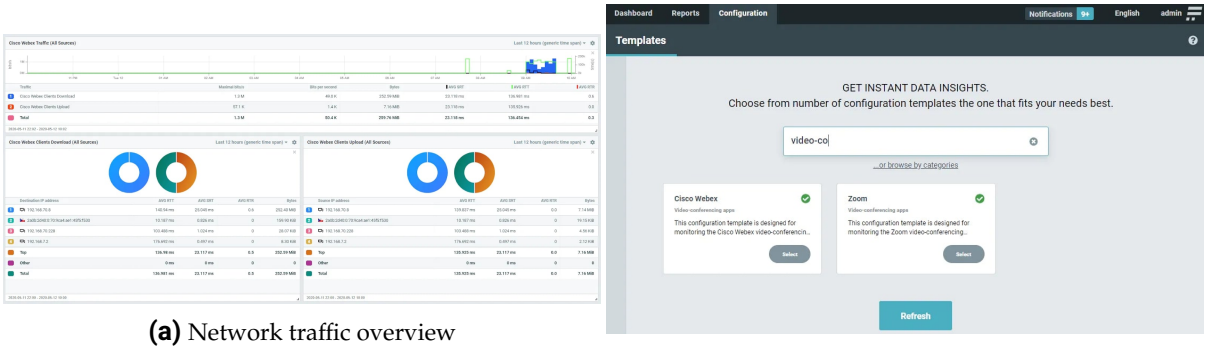
several metrics calculated throughout all hosted meetings on a clean dashboard, also shown in Figure 5.1. These include active users, their geo-distribution, used device types to connect, and also results of custom feedback surveys. What is especially interesting is the possibility of showing some economic metrics regarding Return On Interest (ROI), which includes saved travel costs and reduced emissions. Nevertheless, analogous to Flowmon, Bluejeans is also a very costly software. When hosting meetings, up to 230 € per host per year can be charged for conferences with up to 200 participants. Using webinars for up to 500 participants is priced at 800 € a year. There is the possibility to purchase gateways to Microsoft Teams rooms, which is located at 540 € per year per room with a maximum of 19 rooms. It has to be noted that the command center is only included in the most expensive version of hosting meetings, called *Bluejeans Enterprise*.


**Figure 5.1:** Bluejeans Command Center [Ver22]

### 5.1.2 Flowmon

Another tool in this sector is the software suite called Flowmon [Cor20] presented by Progress Software Corporation, which consists of various modules used for monitoring. It has to be noted that Flowmon not only measures network capabilities but also is strongly positioned in the Application Performance Monitoring Measurement (APM) market. In the specific case

of video conferencing tools, version 11.0 even offers templates for Cisco Webex[3] and Zoom. However, there are some downsides to this tool. Most importantly, with its relatively high price of 15,000 dollars for a lifetime license, including one year of support for the network operations solution, choosing Flowmon is not the solution for everyone. Second, there is also no template for one of the biggest competitor of Webex and Zoom, Microsoft Teams[4], available.




**(a)** Network traffic overview

**(b)** Webex and Zoom templates

**Figure 5.2:** Overview of Progress Flowmon [Cor20]

### 5.1.3 Read.ai

The only free alternative in this comparison is Read.ai [AI22]. Usable as a web app and embedded for other platforms, the tool joins into the meeting call and analyzes all sorts of interactions. In contrast to the other software, it focuses more on the quantitative evaluation of the meeting itself. With a technology-driven approach, a meeting score is calculated based on audio and video engagement. Also, sentiment and overall engagement are calculated as a metric, as seen in Figure 5.3. Remembering Section 2.2, the evolution of these scores is displayed via a time series. One might worry about privacy when this tool uses all transmitted data, but the software claims to delete all data post-meeting and only keep the calculated metrics. There is also the option to opt out during a meeting. If someone writes "opt-out" in the chat, then Read.ai will remove itself and all collected material. A significant advantage over the other two tools is the possibility of being used as an embedded version in Zoom, Webex, and Teams.

### 5.1.4 ZoomSense

However, there is also interesting related work from a research point of view. Bartindale et al. [Bar+21] presented their work with the tool ZoomSense, a scaleable infrastructure that

[3] `https://www.webex.com` (accessed: September 2022)
[4] `https://www.microsoft.com/microsoft-teams/group-chat-software` (accessed: September 2022)

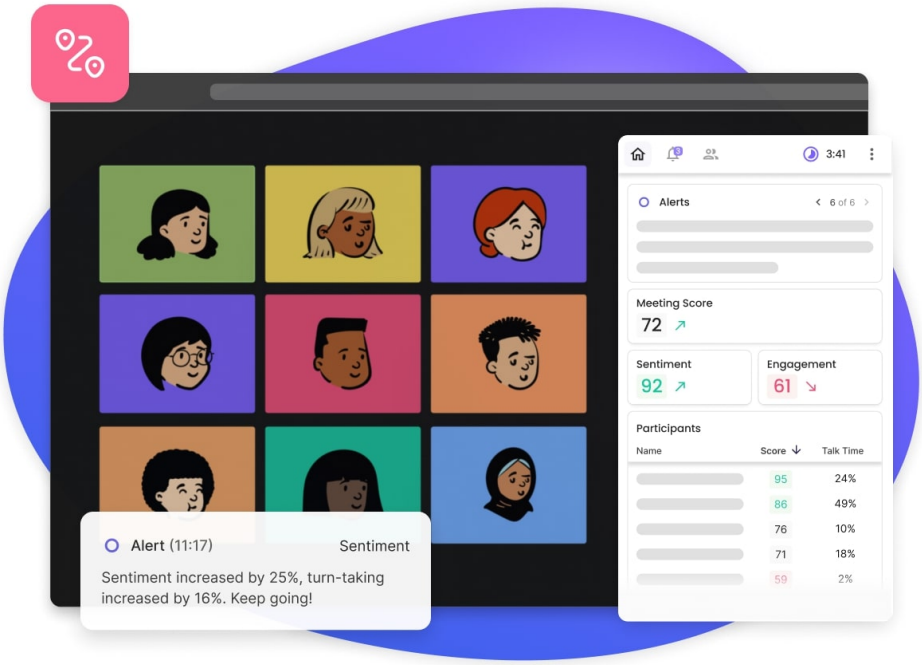



**Figure 5.3:** Meeting Navigator of Read.ai [AI22]

augments Zoom. Its scalability is mainly given through its structure, as ZoomSensors are built as serverless Firebase functions that negate the need to employ a set of Virtual Machines (VMs) for the tool to function. However, their preferred setup is of at least one VM containing a containerized NodeJS scheduler and a set of at last one ZoomSensors, which attend in a single meeting or breakout room. With their connection to the Zoom SDK and the backend, consisting of a Firebase Real-time Database (RTDB), it is possible to observe all actions inside the meeting. With Firebase also communicating with the Zoom API, the attendees can also engage in text dialogue, allowing them to retrieve statistics via chat and stop data collection simply by telling the sensor to "go away." In the context of this thesis, the combination of consent, infrastructure, and interaction with people was influential for the remainder of this thesis. Especially for the practical implementation, some ideas were adapted from the work of Bartindale et al. [Bar+21].

None of the commercial options, except Read.ai, offers automated data collection from software APIs directly to summarize this section. The knowledge gained from ZoomSense is an excellent first step in the right direction. Furthermore, routine collection jobs and introducing custom metrics to fit the users' needs are not possible either. This lack is a significant drawback, as it does not allow for precise statistics that are tailor-made for conferences.
Furthermore, as stated at the beginning of this section, none of the analyzed tools fully comply with our evaluation framework. This divergence can be shown alone in the scope of the

first dimension, where no tool can monitor the identified granularity of conference components. Last but not least, the variety of metrics needed to form a conference fingerprint, as introduced in Section 4.4.11, is neither present nor addable either. Such metrics include environmental and interaction-based ones. Even if one or the other category is present, both of them can not be found in one particular tool. With that in mind, the need for an implementation that fully complies with our evaluation framework is given.

## 5.2 Scope

System design is worthless if it is unknown what goal should even be achieved. Therefore, the requirements should be defined as a first step. Macaulay [Mac12] defines them as follows after the IEEE Standard 610 from 1990:

1. A condition or capacity needed by a user to solve a problem or achieve an objective.
2. A condition or capability must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed documents.
3. A documented representation of a condition or capability as in 1 or 2.

Requirements build the base for the traditional waterfall model. However, this method of software development is only partly used. In combination with the agile project planning throughout the thesis work, the development process in this small scope can be described as partly agile. According to Jørgensen et al., [Jør18], this style is amongst the more effective paradigms for small project sizes. Going full agile would be more implementation overhead regarding the team size.

However, what should be included in this setup? The following functional requirements were defined:

- Possibility to list and search for conferences in a user interface (UI)
- Persist data for conferences over the years
- Query for conferences to compare them
- Show history of the conference (i.e., progression of metrics over the different editions)
- Visualization of content in simple dashboards
- Users should be able to create and additionally insert data for conferences

- The framework should be able to automatically and periodically collect data based on the information from the user input

This list of requirements resulted in a first mock-up of the framework with initial drafts based on known technology used by both team members throughout the university programs and work, seen in Figure 5.4. In the initial draft, the system is accessed via the Conference UI, located in the upper left portion of the mock-up. It should contain table views showing an overview that shows needed information, like, for example, the attendants of the current event, a status flag for reports that have to be collected, and most importantly, all past events of a conference in a separate table. The ability to browse and search for conferences is present as well. The UI is complete with action fields to supply the conferences with missing data by manually ingesting it and adding, editing, and deleting events.
A dashboarding possibility is given in the right portion of the draft to visualize the trend of virtual conferences further. It is invoked by request from the UI, which triggers the framework to produce a JSON file to send to Grafana [Gra22], a tool used by the team before. In the repository of the framework, JSON templates, preferably in parameterized form, are used to replace conference parameters to produce dashboards quickly.
Last but not least, we need to select appropriate storage for the data. For that reason, the storage layer is drafted with a document store, which seems as an excellent first fit.
With additional feedback in hand, two more requirements were added to the list as mentioned earlier:

- Possibility to use and easily implement and register an arbitrary number of data collectors

- Ingest of custom data should be possible

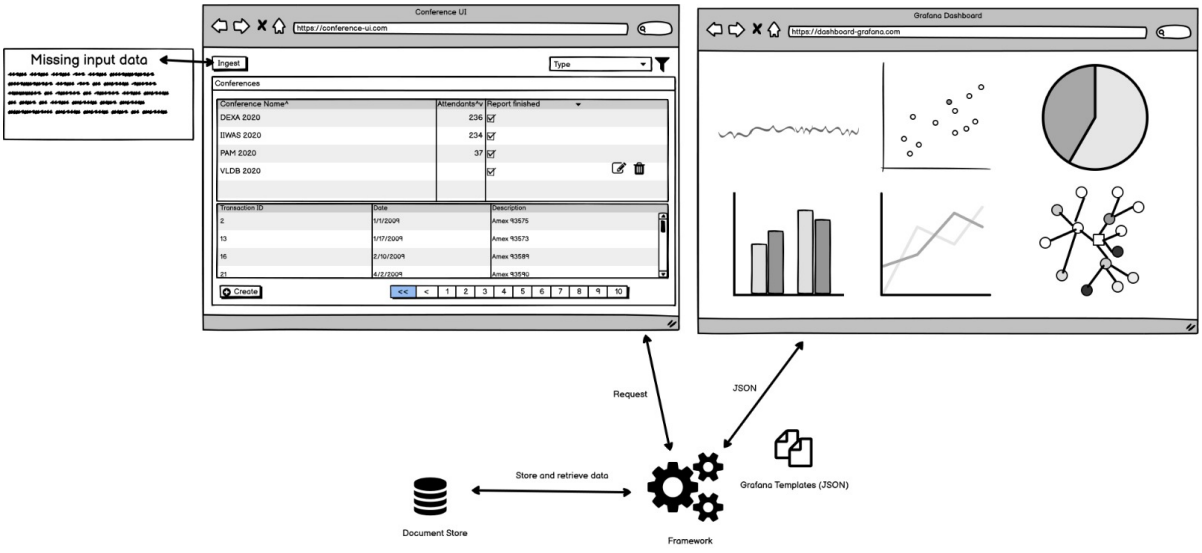



**Figure 5.4:** First mock-up of the proposed framework

However, the mock-up was not fundamentally changed with that additional requirements because they were functional requirements (FRs) towards the framework itself [WB13]. We define therefore as the main use case of this reference implementation a conference information system that supports handling the ingest from different sources and its retrieval and has dashboarding capabilities to visualize trends better.

## 5.3 Selected Metrics

Since the implementation of the reference framework should be suitable for as many conferences as possible, it was necessary to define core metrics for a reference implementation. Apart from the general metrics such as participants, sessions, and costs, as well as information about in which cities and countries the conference has been held over the years, a number of specific metrics should be included. However, with the functionality to add custom metrics, any metrics that are considered relevant for a virtual conference can be included as well. A starting point and guide for selecting suitable metrics can be gained from the refined sample metrics table, see Table 4.4.

Primarily because of the global developments regarding the climate and the pioneering role and exemplary effect of science, we consider it necessary to integrate the metrics of carbon footprint, green innovativeness, and sustainability, see Section 4.3.6. This initiative will monitor the carbon footprint to be kept as small as possible and create an incentive to actively strive for a reduction and use more methods from iteration to iteration to make the conference more environmentally friendly. However, we would like to encourage researchers to select even more green metrics from our refined sample metrics table, see Table 4.4.

As stated in Section 4.3.4, the evaluation criterion of reception is essential for virtual conferences and should be covered with at least one metric. A representative metric concerning virtual conferences is interaction dynamics, which can be tailored specifically to the conference in question. Since, depending on the conference, highly different metrics can be relevant, we have chosen the general term interaction dynamics. For example, concrete reception metrics such as those given in the sample metrics in Table 4.4, e.g., the average number of participants per track and day or an average number of questions per talk, are possible.

## 5.4 Technology Stack

With the requirements and the metrics for this project defined, this section mainly handles the different parts of the framework and which technology fits the specific use case of evaluating virtual conferences in the field of computer science and set of requirements.

### 5.4.1 Programming Language

The main programming language should be wisely considered. For a first overview, a ranking for programming languages is needed. For that reason the TIOBE Index for September 2022[5] is consulted. That list contains community popularity ranking automatically retrieved and calculated based on number of skilled engineers world-wide, courses and third party vendors and is calculated via search engine results. Among the calculation sources are Google, Bing, Yahoo, Amazon, and YouTube. The top four languages from that list contain the following contenders, including their popularity ranking in brackets:

1. Python (15,74 %)
2. C (13,96 %)
3. Java (11,72 %)
4. C++ (9,76 %)

Only programming languages with a score of at least five percent were considered. First of all, due to no experience in the whole team with C++, we exclude that from our argumentation. A comparison is needed between C and Java when considering the available options. Based on the findings from Prechelt [Pre00], C is much more efficient in terms of memory and runtime complexity. Although the garbage collection mechanics of Java have been advanced since this publication when used unwisely, many unwanted objects of variables and references can be created, thus leading to high memory churn [Kho+20]. However, on the other side, C programs do not allow for accessible design. Even with the existence of concepts in C like modularity[6], design for simple and extendable Application Programming Interfaces (APIs) is not possible. For that reason alone, C is neglected from that choice because it contradicts the defined requirements.
After that, the decision has to be made between Python and Java. Both languages have their strengths and weaknesses based on the comparative analysis of both languages by Khoirom et al. [Kho+20]. One of the more essential facts is the limitation of database access layers. In that regard, Java should be explicitly favored by JDBC (Java DataBase Connectivity) and

[5] `https://www.tiobe.com/tiobe-index/` (accessed: September 2022)
[6] `https://www.geeksforgeeks.org/modular-approach-in-programming/` (accessed: September 2022)

ODBC (Open DataBase Connectivity). With the Global Interpreter Lock (GIL) of Python, multi-threading is an impossible task. Further, Python's simplicity, especially with its focus on dynamic typing, is a vast disadvantage when embedding into stricter languages, like Java.

With this comparison of both contenders, the presence of concepts like JDBC and multi-threading are crucial to fulfilling the given requirements [Kho+20]. With these points, some external software with their libraries has to be used. Another factor to be considered is the preference of the team to use Java, as we are most experienced developing in that specific language. Therefore, the final decision for a programming language for this framework is *Java*.

### 5.4.2 Server Backend

In this section, we want to dive deeper in the question: What do we need as a foundation to successfully build our architecture on? We need a Java-based framework that allows for an easy setup and is lightweight in terms of third-party dependencies. In an optimal case, standard functionality for communication like transaction management, user authentication, and data exchange with the frontend is handled to reduce boilerplate code further.

#### Spring Boot

After researching our needs for a fitting framework, we found a promising solution. Therefore, we use the Java-based Spring platform[7] includes several modules such as Spring Security, Spring JDBC, and Spring Test, and offers features such as templates, and dependency injection, which provides required Java objects in the context of a class that depends on these objects, and inversion of control, which manages the lifecycle of these objects. In addition, boilerplate code is reduced because proven standard functions are already integrated and can be used directly, e.g., data access and transaction management[8].
Spring Boot[9] is an extension for the Spring platform, which facilitates development in many respects. Unlike the Spring platform it is built on, Spring Boot requires less configuration, especially for the third-party libraries used, and is, therefore, easier to set up. Spring Boot combines the benefits of the Spring framework but provides standard configurations that have been proven in development. The Spring Initializr project configurator[10] can be used to

[7] `https://spring.io/` (accessed: September 2022)
[8] `https://spring.io/why-spring` (accessed: September 2022)
[9] `https://spring.io/projects/spring-boot` (accessed: September 2022)
[10] `https://start.spring.io/` (accessed: September 2022)

quickly select the right libraries and include them correctly in the configuration files. In addition, Spring Boot provides an already configured and embedded Tomcat web server. Another reason for choosing Spring Boot is that third-party libraries are monitored by security experts and only supported if they pass their review of vulnerabilities[11]. Especially for the further development of our reference framework, it was important for us to rely on open-source tools that contribute to better extensibility and flexibility. Therefore, we selected Spring Boot to implement the backend for our reference framework.

#### Spring Security

In addition to Spring Boot, we integrated Spring Security[12] in our project. Spring Security offers protection against common threats, e.g., Cross-Site Request Forgery (CSRF), a cryptography module to encode passwords and key generation, HTTP Strict Transport Security (HSTS), which is effective against Man in the Middle (MITM) attacks, as well as protection and utilities for user authentication.

### 5.4.3 Storage and Data Access Layer

The heart of every data-intensive application is the storage and data access layer. Without the possibility to store and retrieve data at will, every request can only live temporarily. For that reason, DataBase Management Systems (DBMS) exist. Initially, only Relational DBMS (RDBMS) existed. One of their main paradigms was ACID, short for Atomicity, Consistency, Isolation, and Durability [Kle17]. However, with "Big Data," the need for more availability and scalability in distributed systems arose [MK14].
The relational model can not efficiently handle the current needs of large and often unstructured datasets. Thus, the BASE principle of Basically Available, Soft State, and Eventual Consistency has emerged to fill this gap for NoSQL databases [KW19]. It is also associated with reliability, as mentioned earlier requirement category. As an important side note, although directly compared in many cases, ACID vs. BASE compares two different scopes [MK13]. The first principle aims toward modeling transaction-level integrity, whereas the latter considers the system's design as a whole [MK14]. One of the main aspects of our choice of NoSQL systems in the storage area is the proposed flexibility of the framework. With BASE, a high focus on scalability and flexibility is set, thus a perfect fit for our implementation.

However, what about the data and how it is modeled internally? Here, the rationale for design patterns in NoSQL data modeling arises. The specific need for an item, defined via a

[11] `https://spring.io/why-spring` (accessed: September 2022)
[12] `https://docs.spring.io/spring-security`, accessed: September 2022

query, is the driving guidance for the modeling task. In literature, they are called *query access patterns* [RS20].
Such an access pattern is defined as the single element of an application's task to aggregate data from or insert data into the database to fully work [GCR16]. They can comprise read as well as write access patterns [Khu12]. With the requirements defined and the mock-up refined, all discussed in Section 5.2, six access patterns were defined to cover our needs. These are in particular:

- User creates a new account
- User creates a new conference
- User ingests data about conference
- Data collectors ingest collected data about conferences
- User wants to authenticate
- User wants to request data about one or more conferences

In that regard, only four primary entities to store are relevant for this use case. Figure 5.5 shows the complete data model in an UML class diagram. The main element here is the conference, consisting of the title as its identifier. Other important information is the organization and publisher as metadata for a conference. It should be handled in a centralized way to avoid unnecessary data duplication. However, recent cases from iiWAS[13] and DEXA[14] should be used as thought-provoking examples, as the publisher is going to change respectively the organizer changed. For the sake of simplicity, we stay at the original design decision to keep these metadata fields at the original entity. Nevertheless, this decision has to be monitored and reconsidered for future work. A ConferenceEdition is a sub-entity of Conference and contains all information needed to capture one singular event. Also, the metrics selected in Section 5.3 and modeled after the sample metrics found in Section 4.4.10 can be found here. With the basic conference structure handled, a model for extending the basic set of metrics with additional metrics provides more flexibility. With the identifiers, both in numeric and text form, and the datapoint and a reference to the conference edition stored, a time series, defined in Section 2.2 can be constructed. In this case, the reference to the edition acts as a pseudo timestamp. To enable automatic data collection, as discussed in Section 4.4.6, ingest configurations for additional metrics are stored. To provide further flexibility, ingest types can be defined with an arbitrary number of configuration parameters used for retrieving data from a source. For example, a REST endpoint and a path for parsing the correct value can be stored in the parameters.

[13] `http://www.iiwas.org/` (accessed: September 2022)
[14] `http://www.dexa.org/` (accessed: September 2022)

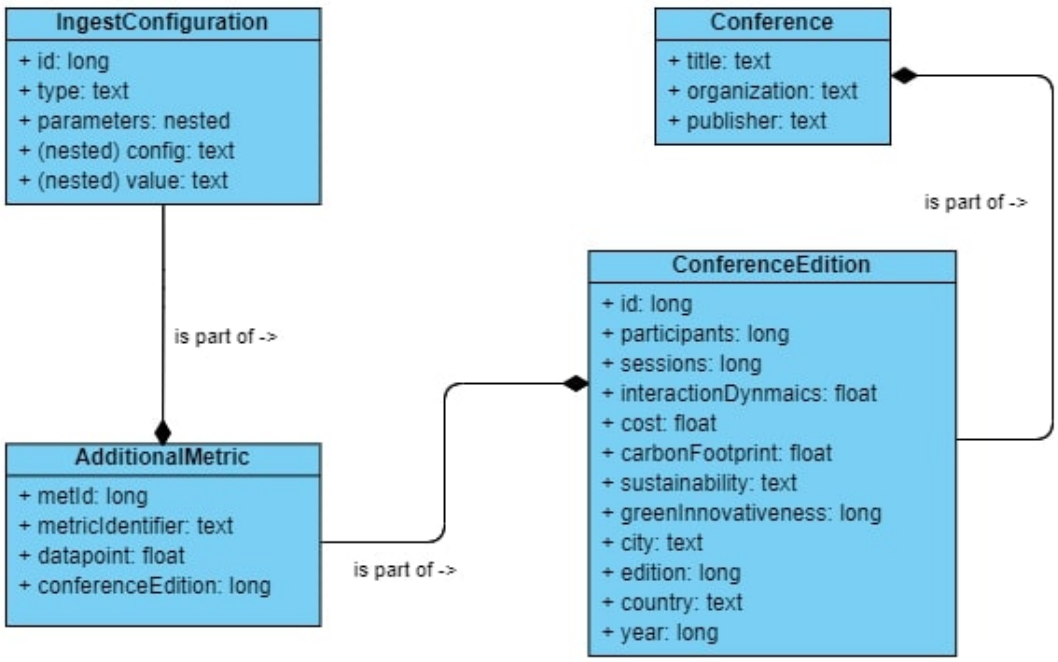



**Figure 5.5:** Data model for handling the given access patterns

With the data structure properly handled, the main part of this project is the specific database to be used. As seen in Figure 5.5, the IngestConfiguration class has two nested fields found under a field called *parameters*. This nested type resembles the possibility of ingesting and storing an unknown amount and variety of data as additional information. In this case, an arbitrary number of pairs, consisting of a config-key and a value, is ingestable. Data of unknown size and structure is not easily storable in relational databases [Kle17]. Therefore, only NoSQL systems are left in this decision-making process. When examining the structure of the data model, a conference can be seen as a collection of documents, and an edition is one document of this collection. For that reason, document stores seem to be a good fit. Search engines with the capability of storing documents are also beneficial. Comparing the top choices for each category, DB Engines by solidIT[15] suggests MongoDB [Mon22] as the most popular document store and ElasticSearch as the best choice for search engines [Ela22]. When comparing both choices, in light of this application, which solely has access patterns for ingesting and creating documents for conferences, the main load will lie on the read patterns to search for data. Examples of that could be the conferences with the lowest environmental impact in European cities. Therefore, Elasticsearch will be the main choice in that regard.
This selection has another important advantage: less need for more external libraries. The magic term in that context is called the *Elastic-Stack*, formerly known as the ELK-Stack[16]. This tech stack consists of three additional tools besides ElasticSearch: Kibana, Logstash, and Beats.

[15] `https://db-engines.com/en/ranking` (accessed: September 2022)
[16] `https://www.elastic.co/what-is/elk-stack` (accessed: September 2022)

While Beats and Logstash are not relevant for this specific use case, Kibana is even more important, as it allows for extended dashboarding capabilities for Elastic. But also solutions of other companies are supported, like DataDog[17] and Dynatrace[18]. Kibana specifically will be discussed in Section 5.4.4.

### 5.4.4 User Interface

This section describes the two major frontend parts, defined in the mock-up in Section 5.2, further: (i) the web interface, and (ii) the dashboarding component.

#### Web Interface

For our master thesis project, it is necessary to enable the user to display data about the conferences clearly and to be able to add data easily. Furthermore, it should be possible to search within the listed conferences for the desired information and apply filters and advanced sorting. Furthermore, the interface must be easy to use and navigate. The interface itself should be easily integrated into the existing technology stack. Therefore a web interface should be created which fulfills these criteria.

Since we are using Spring Boot as the backend and Java framework, choosing a complementary technology that can be used in conjunction with it makes sense. Besides the popular frontend frameworks like Angular[19], React[20] and Vue[21], which would undoubtedly all be a suitable choice, it is also possible to use the combination of Thymeleaf[22] and Bootstrap[23], which can easily and seamlessly be integrated into our technology stack. Bootstrap is the JavaScript framework with the second highest market share, being utilized on 26.6 % of all websites whose frontend is known, according to W3Techs[24].

Bootstrap is an open-source frontend toolkit primarily based on CSS and JavaScript that includes a pre-configured grid system and ready-made styling components that can be used like building blocks for projects. It is also possible to integrate JavaScript plugins and extend Bootstrap. Furthermore, Bootstrap is geared towards mobile-first and responsive projects. Another reason for using Bootstrap is that most components already comply with WCAG 2.0 and other accessibility standards. For example, interactive components such as custom

[17] `https://www.datadoghq.com` (accessed: September 2022)
[18] `https://www.dynatrace.com` (accessed: September 2022)
[19] `https://angular.io/`, accessed: September 2022
[20] `https://reactjs.org/`, accessed: September 2022
[21] `https://vuejs.org/`, accessed: September 2022
[22] `https://www.thymeleaf.org/`, accessed: September 2022
[23] `https://getbootstrap.com/`, accessed: September 2022
[24] `https://w3techs.com/technologies/details/js-bootstrap`, accessed: September 2022

tooltips, dropdown menus, and modal dialogs are designed to be accessible to keyboard, mouse, and touch[25]. Bootstrap is typically installed with a package manager, or the compiled CSS and JS files are included via a CDN.

Thymeleaf is a Java templating engine mainly used to generate HTML views. Thymeleaf can also be used as a general-purpose templating engine. For example, a Thymeleaf template can consist of an HTML page and Thymeleaf expressions. It is also possible to load dynamic content with Thymeleaf expressions. Furthermore, Thymeleaf expressions can access Java code and objects and Spring beans. A web browser connects to the Spring controller that interacts with the Spring model and accesses the Thymeleaf template. The processed result in an HTML page is returned to the web browser. So when working with a web app, Thymeleaf is processed on the server, not on the client side. Thymeleaf also provides loops and conditions and integrates CSS and JavaScript. With the Spring Initializr[26], we can add the appropriate dependencies and easily include and pre-configure Thymeleaf.

With this combination, we not only get a great deal of flexibility and ease of use, but we also have greater control and direct access to Java objects and Spring beans. This characteristic makes it easier to link the backend and front end and access the benefits of Spring Boot, which handles many deployment and configuration tasks.

### Dashboarding

Another requirement was the possibility of visualizing data of conferences on a webpage and showing the conferences' history over time, allowing for comparison of conference editions over the years, as it was needed for answering RQ3. The latter is easily handled because such metrics, observed over time, can be seen and visualized as time series. This concept was also discussed in Section 2.2.
Visualization is more complex than that. To achieve this, also much technology is available to developers. Remembering the draft in Section 5.2, Grafana was used as a possibility to do so [Gra22]. Grafana is a free-to-use tool allows many different data sources to connect for visualization. ElasticSearch is even supported via built-in connectivity. Also, present dashboards could be persisted in Elastic, as Grafana saves their templates as JSON documents. However, one significant shortcoming is the embeddability of the UI, as the resulting dashboards would be on different pages respectively and would need Grafana authentication. All in all, this would be a non-seamless process for the user.

[25] `https://getbootstrap.com/docs/4.0/getting-started/accessibility/`, accessed: September 2022
[26] `https://start.spring.io/`, accessed: September 2022

Therefore, Kibana, a tool from the Elastic-Stack was also considered[27] and was favored over Grafana. One reason for the selection was that it was already included as part of the stack and did not need extra setup. Furthermore, the strong coupling to ElasticSearch [Ela22] allowed for more query functionality. Many setting possibilities also allow for clean dashboards, as seen in Figure 5.6. Last but not least, dashboards could also be embedded into other pages.

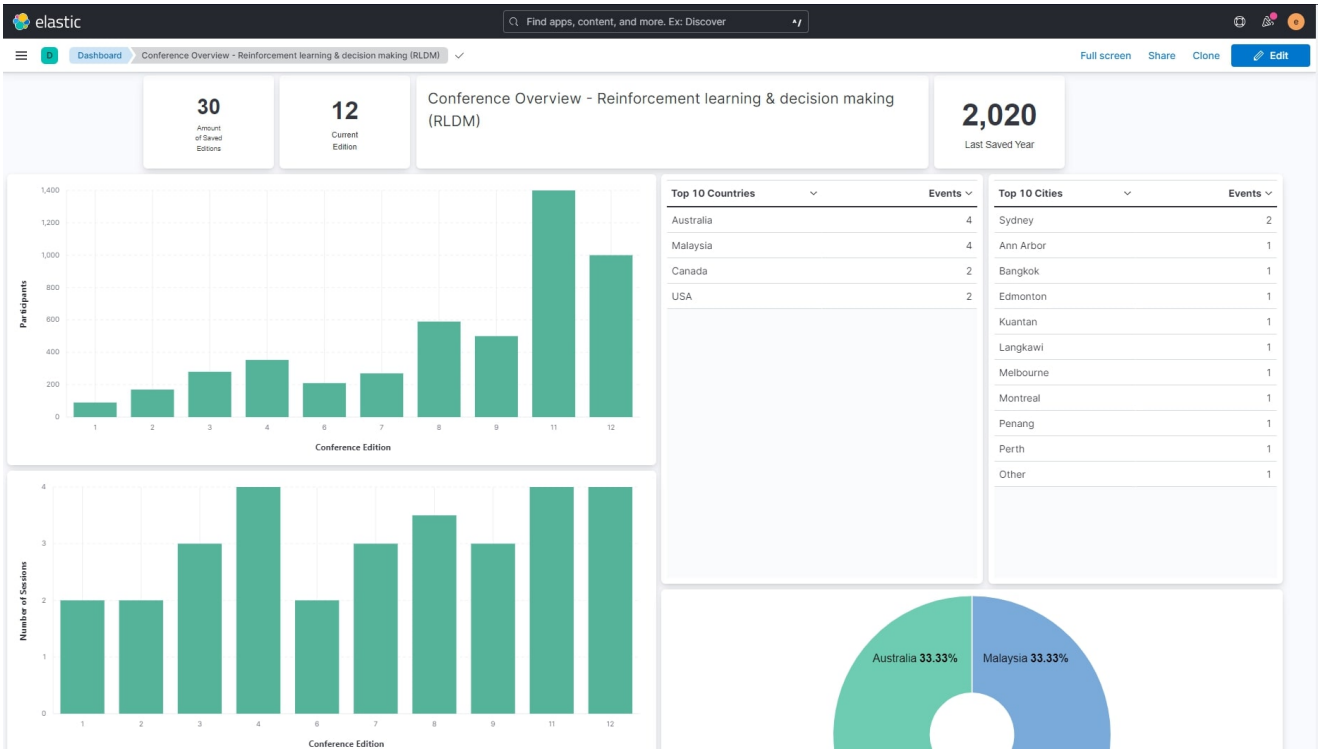



**Figure 5.6:** Sample Dashboard for a Conference, made with Kibana [Ela22]

## 5.5 Test Datasets

In relation to the aforementioned selection of metrics used for an initial implementation, selected data about scientific conferences were used. It has to be noted that at the time of writing this master thesis, suitable data for evaluating virtual conferences is scarce, especially in terms of the newly designed evaluation criteria and their according metrics. For this reason, we had to craft a data set from several sources, according to the principles of data collection that we established in Section 4.4.3. This data was also condensed into a set of Comma Separated Value (CSV) files to have accurate test data, which can be used for demo purposes. Additional data was sourced from the homepages of the selected conferences. The primary source of data is the Improving Conferences Github page of eLife Community Ambassadors[28]. The sam-

[27] `https://www.elastic.co/kibana/` (accessed: September 2022)
[28] `https://elifeambassadors.github.io/improving-conferences/` (accessed: September 2022)

ple dataset is divided into four files, analogous to the data model, discussed in Section 5.4.3, consisting of the CSV-files Conference, ConferenceEdition, AdditionalMetric, and IngestConfiguration. In total, the following amount of data can be found:

- 3 conferences (ICCAE[29], RLDM[30], ICSCA[31])
- 4-5 editions per conference (total of 13)
- 2 metrics per edition (total of 26)
- 1 ingest configuration per metric (total of 26)

A sample line per file could look like the data from Tables 5.1, 5.2, 5.3, and 5.4

| title | organization | publisher |
|---|---|---|
| International Conference on Computer and Automation Engineering (ICCAE) | ICCAE | ACM |

**Table 5.1:** Excerpt of the Conference sample data file

| conferenceId | year | edition | participants |
|---|---|---|---|
| 1 | 2020 | 12 | 1000 |
| | | | |
| sessions | interactionDynamics | cost | carbonFootprint |
| 4 | 0,6 | 400000 | 2500 |
| | | | |
| sustainability | greenInnovativeness | city | country |
| 0 | 0 | Sydney | Australia |

**Table 5.2:** Excerpt of the Conference Edition sample data file

| metricIdentifier | data point | conferenceEdition |
|---|---|---|
| Acceptance Rate | 20 | 1 |

**Table 5.3:** Excerpt of the Additional Metrics sample data file

This setup only covers a half-automatic approach since the data was structured into CSV files beforehand. In case a data repository can be contacted via REST or another query language, the code could be adapted to automatically parse and ingest new conference data in case there is no entry in our system.

[29] `http://www.iccae.org` (accessed: September 2022)
[30] `https://rldm.org` (accessed: Sept. 2022)
[31] `http://www.icsca.org` (accessed: September 2022)

| metricId | type |
|---|---|
| 1 | manual |

**Table 5.4:** Excerpt of the Ingest Configuration sample data file

## 5.6 Implementation of System Architecture

This section shows more insight into the design and implementation of the reference architecture. The architecture ran through a holistic redesign process, to simplify the overall complexity of the system. This process was possible thanks to thought-provoking assumptions listed in Professor Ousterhout's *A Philosophy of Software Design* [Ous18] to manage complexity in a software system.

### 5.6.1 Choice of System Architecture

For the system architecture of our practical project, we need to consider which one best meets our requirements. Since the evaluation tool is extensible and fed from multiple heterogeneous data sources, it makes sense to choose a microservices architecture. However, utilizing microservices introduces its own set of problems. García [Gar20] recommends the small monolith approach. It features compartmentalization of services similar to microservices but with a single codebase in the same programming language that facilitates the application's development and deployment. On the one hand, especially for small teams, García [Gar20] emphasizes that the technical complexity of microservices should not be underestimated and that they might add a layer of unnecessary difficulty to development. On the other hand, monolithic architectures are mostly organically grown system architectures extended repeatedly over the years and exhibit tight coupling [Roc22]. The more this type of architecture grows, the more difficult it is to maintain. Personal services or functions are integrated into the whole and cannot be used separately. They cannot be easily reused in this way for other projects [Roc22]. There are no clear boundaries between services, and further development can be difficult as the codebase gradually becomes more complex and harder to read over the years [Roc22]. However, a small monolith architecture can be easily transformed into a microservices architecture when the need arises [Gar20]. Therefore, this balanced approach of a small monolith architecture seems more feasible.

### 5.6.2 Structure Diagram

The architecture of the reference framework is divided into different packages that combine main functionalities, see Figure 1. This structure is intended to support the change from a small monolith architecture to a microservices architecture through its clear structure

and the subdivision into functional units, as detailed in Section 5.6.1. The main package `masterthesis.conferences` contains the executable `ConferencesApplication` and the corresponding `ErrorFilter`. With the aid of the `ConferencesApplication`, the server can be initialized, and the Spring application can be executed. In the `data` package, all those classes are bundled, which concern the data model, data transfer objects (DTOs), or metrics. For this reason, `data` contains the subpackages `metrics`, `model`, and `util`. The `server` package consists of classes representing the application's control, display, and storage logic. Accordingly, `server` contains the subpackages `controller`, `dashboarding`, and `rest.service`. Figure 5.7 visualizes this simple structure by showing a simplified class diagram that neglects all fields and methods. DTO classes and implementations of base classes and interfaces are also omitted for increased readability. The diagram focuses mainly on the structure of the architecture. The full diagram can be found in Appendix Section 2.

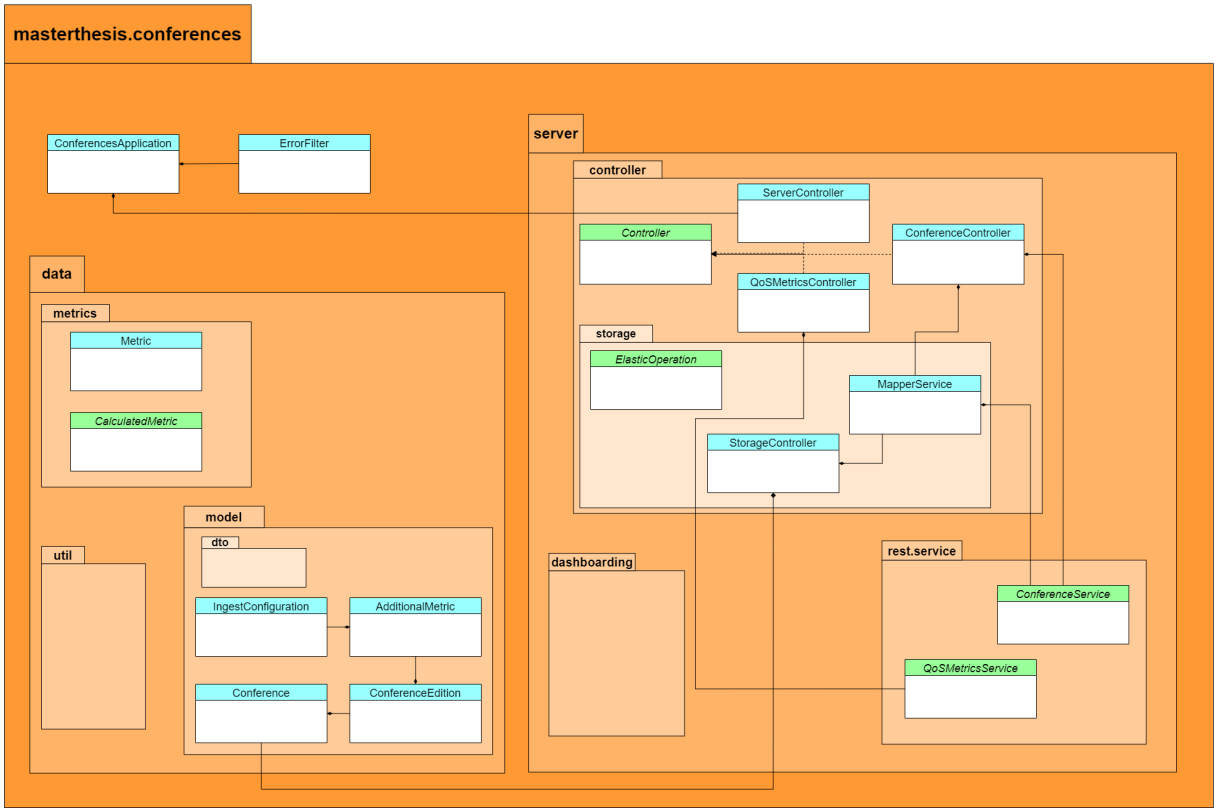



**Figure 5.7:** Simple structure of classes

### 5.6.3 Core Features and Functionalities

We will first start with an insight into the reference framework's essential functions before describing the managing and monitoring processes in more detail.

The implementation of the reference framework allows managing and monitoring conferences to improve them from edition to edition. In the overview, see Figure 5.8, all recorded conferences can be displayed and managed. All the standard operations, such as creating, deleting, editing, and listing conferences, are accessible from this overview. Additionally, the bar with the buttons on the right-hand side offers the selection of displayed fields, advanced search, and sorting functionalities. The listed conferences can be sorted in ascending or descending order according to the data in their fields. It can also be selected which fields should be displayed so that the overview can be limited to the most relevant fields. The advanced search offers the functionality to search for a term in specific fields. Above the listing of conferences, a full-text search field is placed, allowing for search across all conferences. The design is responsive and can be resized to fit different screen sizes. In the overview, it is also possible to generate and display a dashboard for a conference.

Another important consideration is that the reference framework has been designed with further improving accessibility in mind. During the development process, the components of the UI were evaluated with WAVE, the Web Accessibility Evaluation Tool [32], and adapted if necessary. In addition, whenever possible, accessibility measures were taken from the outset. It must be noted that Thymeleaf produces Accessibility issues, which are detected by WAVE and could not be resolved by hand.

After this overview, we consider the process of management and monitoring virtual conferences next. First, a conference is created with the basic data such as title, publisher, and organization; see Figure 5.9. Afterward, editions can be added to the conference, which include, for example, the conference's location, the country and year in which it was held, and the number of participants and sessions, see Figure 5.10. Any number of conference editions can be added. Additional metrics can be added to each conference edition; see Figure 5.11. These metrics can be defined arbitrarily. Only data points, metric identifiers, and the ingesting type must be specified. Parameters necessary for the read-in can be created in an ingest configuration; see Figure 5.12. With the ingest type, it is possible to read the preconfigured type `zoom` for reading sample metrics via a mocked-up `Zoom API`. The reference framework also includes a cron job for automatically reading in sample metrics periodically from this mocked `Zoom API`.

After creating the conference and at least one edition with as many additional metrics as needed, we can create a dashboard for it in the overview, see Figure 5.8. This sample dashboard has been created for the conference Reinforcement learning & decision making (RLDM), based on real-world data. Before creating the dashboard, we need to select the additional metrics that should be displayed; see Figure 5.13. On this screen, the panel title, panel layout, e.g., bar, line, or metric, as well as the metric operation, e.g., median, max, min, average, and count, can be specified. After this process is completed, the redirection to the generated

[32] `https://wave.webaim.org/`, accessed: September 2022

Kibana dashboard takes place; see Figure 5.6. Each tile contains different information, from simple statistics to histograms of the number of participants or sessions per conference edition and pie charts. The tiles can be selected and arranged differently. As a unique feature, the additional metrics created with custom tiles can also be added; see Figure 5.14. In this example, the median participant satisfaction of the RLDM conference is displayed as a histogram over all available conference editions.

Since the relevant metrics for a conference can be added, the framework can be flexibly tailored to the specific conference. This way, metrics for each conference holding can be tracked over the years and analyzed with visual support. The dashboard also provides an interactive report of all the recorded details of a conference and, last but not least, can be used for documenting and preserving knowledge.

In summary, the reference framework offers the following core features and functionalities:

- Flexible, expandable framework
- Components designed with accessibility in mind
- Providing an overview of all registered conferences
- Managing conference data (creating, deleting, editing)
- Adding custom metrics
- Generating dashboards with tiles for custom metrics
- Full-text search over all registered conferences
- Advanced search functionalities
- Sorting based on available fields (ascending and descending)
- A cron job to automatically ingest sample metrics from a mocked `Zoom API`

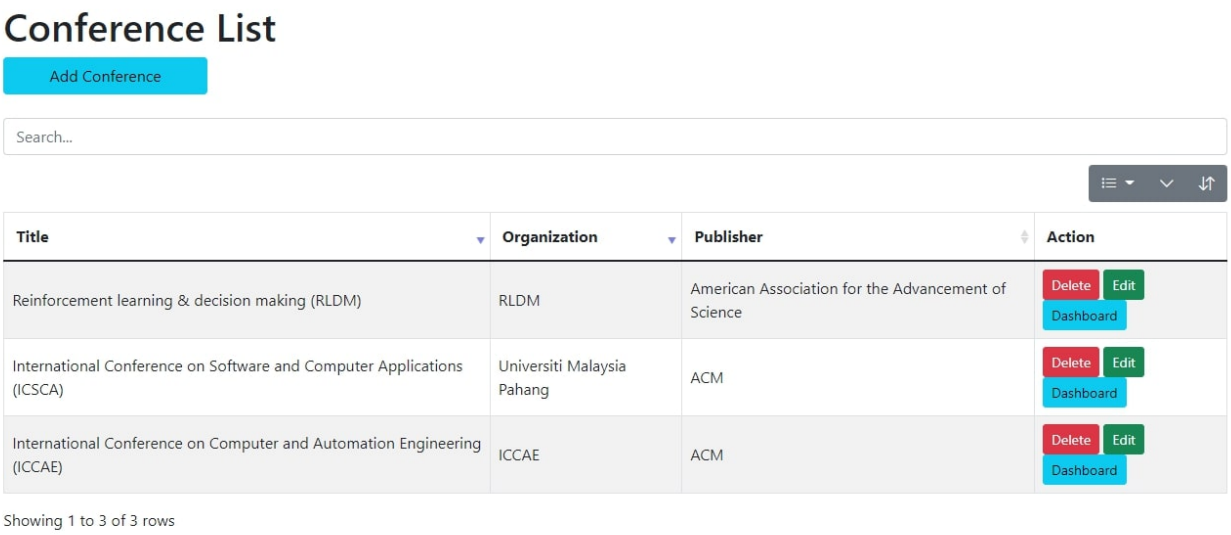



**Figure 5.8:** Overview of Conference Interface

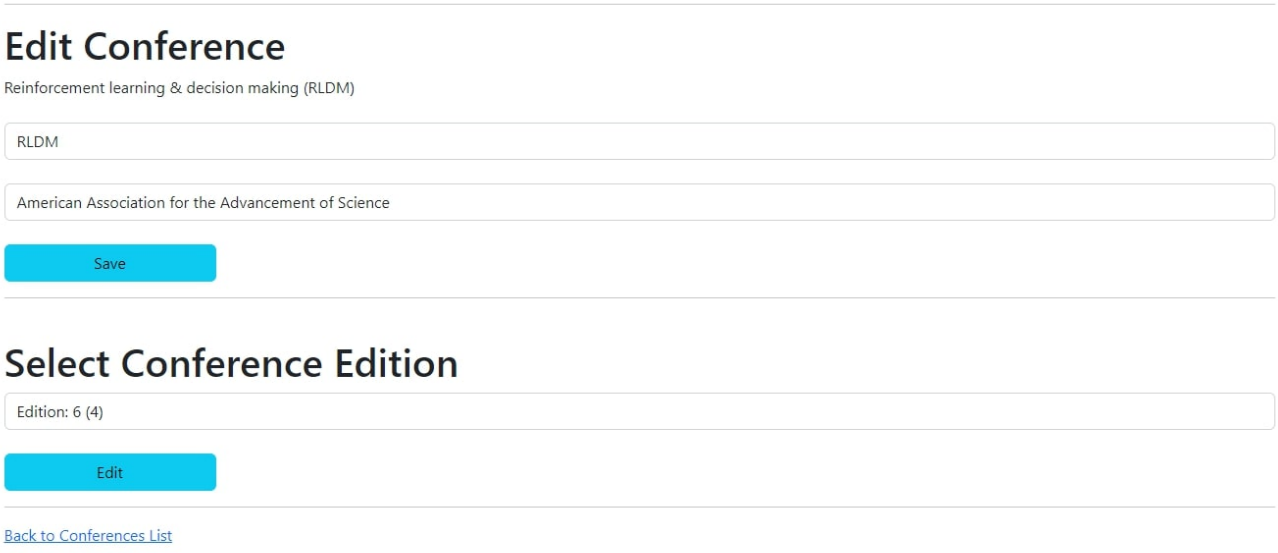



**Figure 5.9:** Edit Conferences Screen

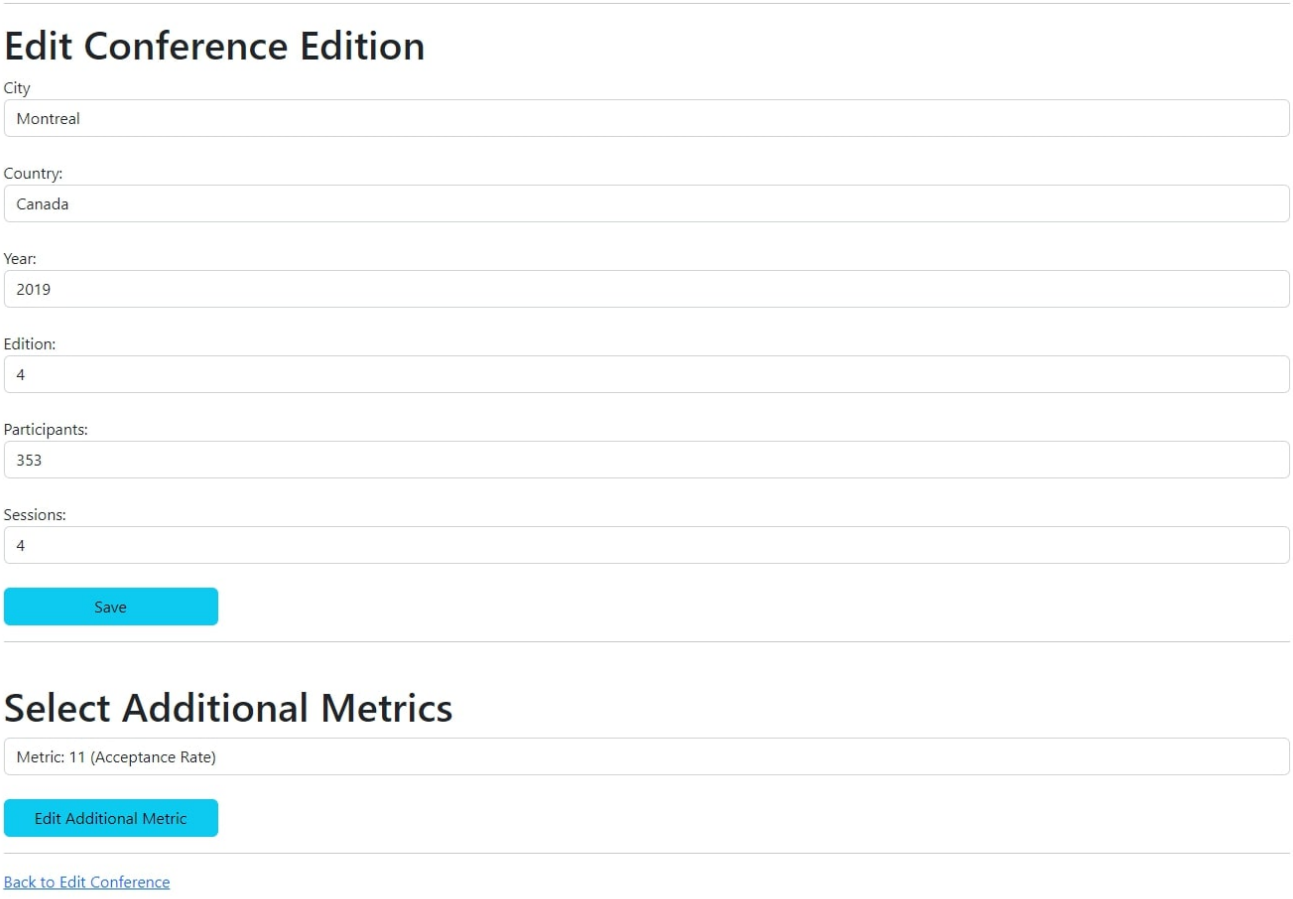



**Figure 5.10:** Edit Conference Edition Screen

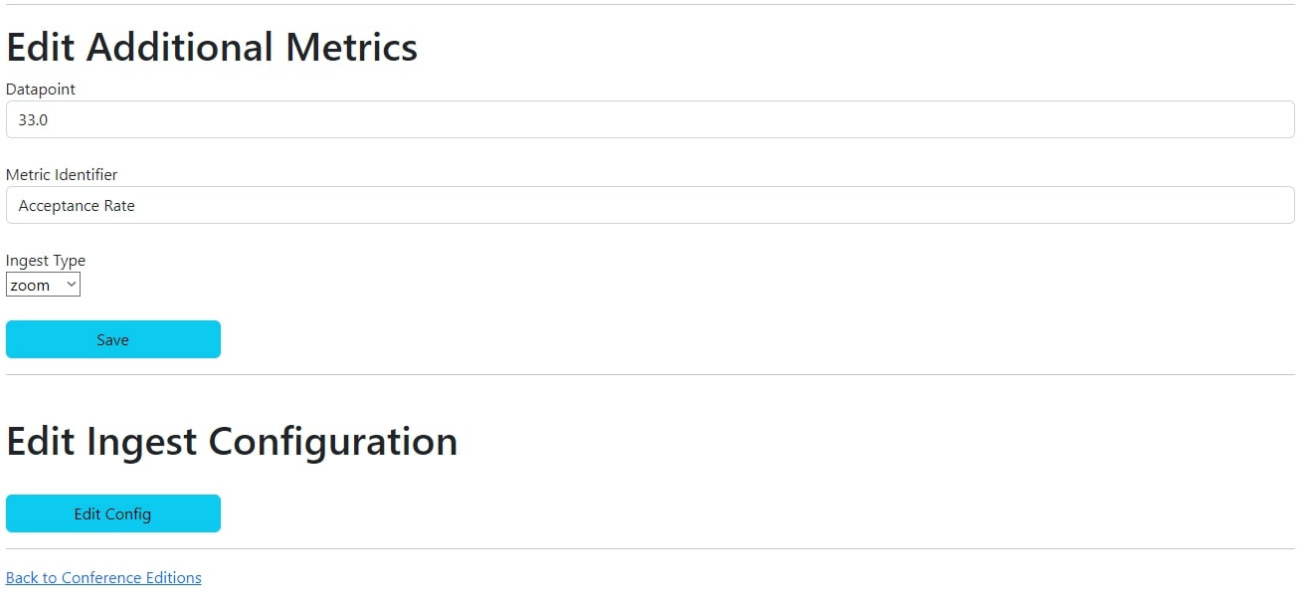



**Figure 5.11:** Edit Metric Screen

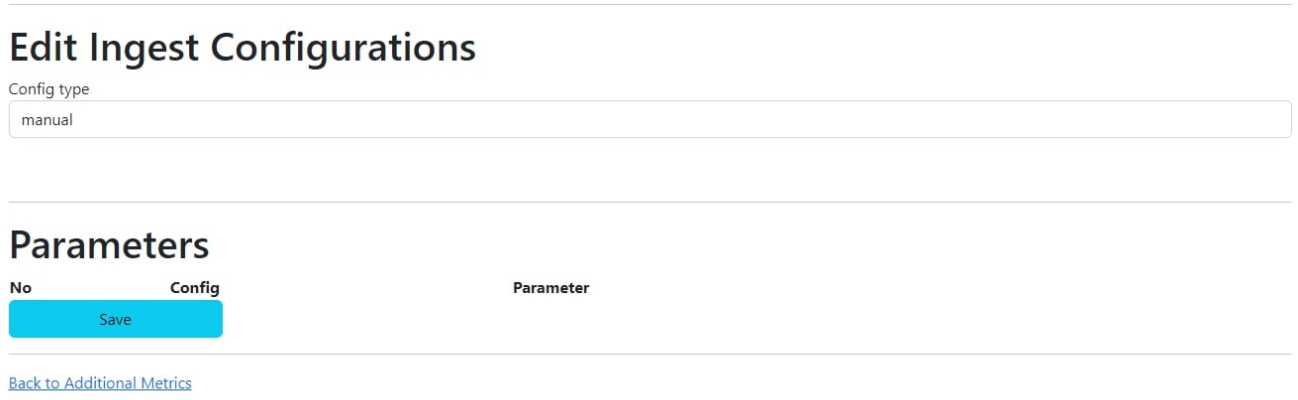



**Figure 5.12:** Edit Configuration Screen

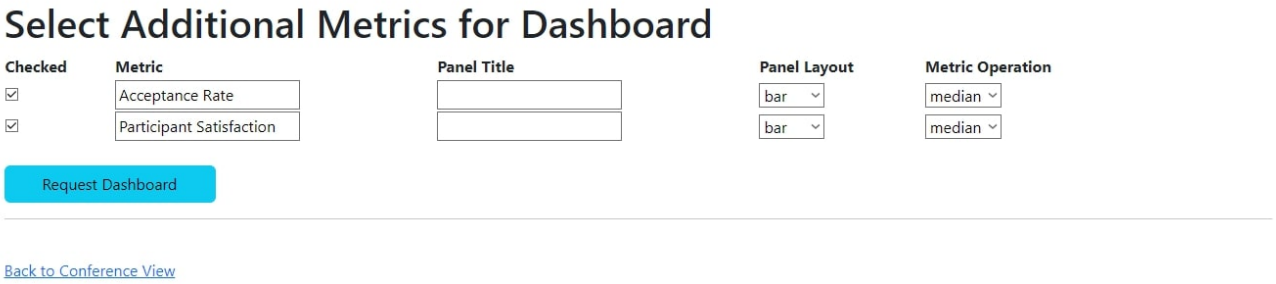



**Figure 5.13:** Select Additional Metrics for Dashboard Screen

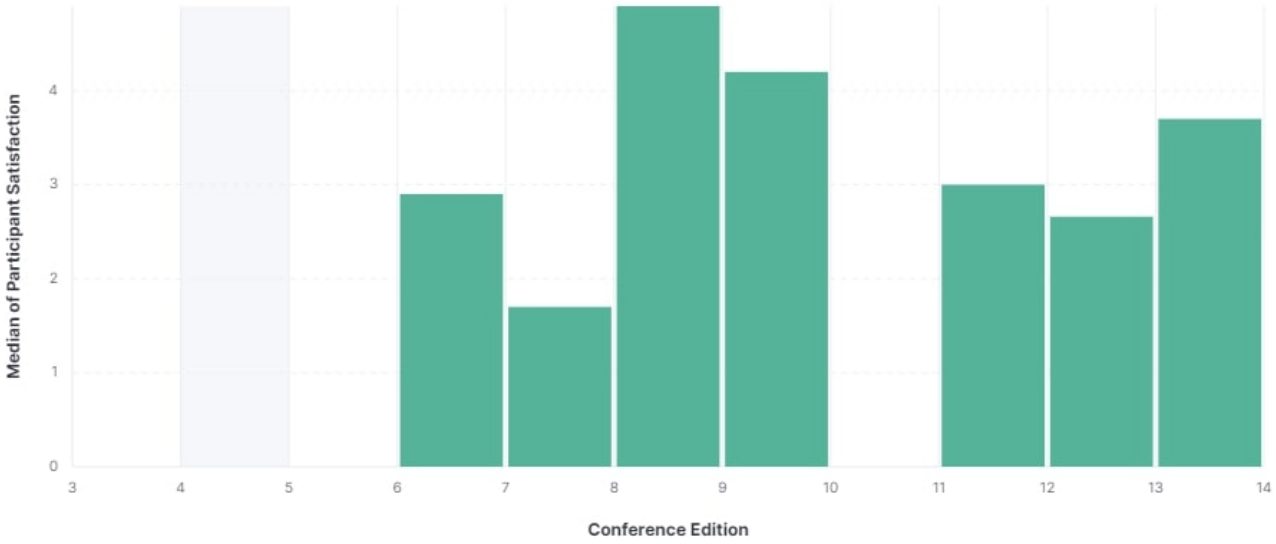



**Figure 5.14:** Custom Metric for a Conference

### 5.6.4 Software Evaluation

For our implementation of the reference framework, we set the following software evaluation goals:

- The project should be easy to maintain
- Classes should be reusable with a low number of dependencies
- The complexity of the project should not be too high in order to facilitate understanding

It is necessary to evaluate the software to ensure that the implementation of our reference framework meets objective standards. For this purpose, we use the object-oriented metrics of Chidamber & Kemerer, which are used in many evaluation tools and studies and have been proven informative and reliable [MNM20]. Their reference set of metrics consists of the following metrics [CK94]:

1. **Coupling between Object Classes (CBO)**: CBO measures the coupling between classes, i.e., how methods and variables of a class are accessed in another class [CK94]. The lower the CBO, the more reusable a class is [CK94]. A high CBO is indicative of higher fault-proneness, a more difficult maintenance process, and the necessity of more rigorous testing [CK94]. We aim for a CBO lower than nine, based on the reference data [CK94; BBM96].

2. **Depth of Inheritance Tree (DIT)**: DIT measures the level of the hierarchy tree a class is located. A high DIT has been associated with increased complexity and higher fault-proneness [BBM96]. The average DIT should be as low as possible, e.g., below 5-10, depending on the size of the project [CK94]. In our case, we define a limit $< 8$.

3. **Lack of Cohesion in Methods (LCOM)**: LCOM measures how disparate a class is, i.e., if it can be split into separate classes, because of methods that access disjoint sets of instance variables [CK94]. A high LCOM indicates fault-proneness and that the software should be redesigned to have more cohesive classes [CK94]. Our project aimed for an LCOM value below ten since we favor cohesive and reusable classes.

4. **Number of Children (NOC)**: NOC accounts for the number of immediate child classes, which means that the breadth of a class is measured. This metric is strongly dependent on the size of the project. A high NOC in conjunction with a high WMC signifies high complexity and the need for redesigning the software architecture. The larger the NOC value, the harder it is to detect faults, which is why classes with a larger NOC appear less fault-prone [BBM96]. An ideal NOC is typically lower than 1 [BBM96].

5. **Response for a Class (RFC)**: RFC measures a class's response set of methods that are typically executed when a response is triggered. RFC also includes methods from outside the respective class, but only the first level of calls [CK94]. Classes with a high RFC are generally more complex [BBM96]. As RFC strongly depends on the project size, an optimal value must be explicitly selected for a project. For our framework, we aimed for an RFC below 50, which is less than half of the maximum value.

6. **Weighted Methods per Class (WMC)**: WMC measures the method count for a class. A high WMC signifies large classes with many methods, which can disadvantage maintenance and reusability. On average, the WMC should be lower than 20 to 50, depending on the size of the project [BBM96]. We therefore define a limit of $< 30$.

A tool for this evaluation we use the plugin MetricsReloaded for IntelliJ[33], which is recommended by Molnar et al. [MNM20]. During the development of our implementation of the reference framework, we utilized the results of the plugin to check if we were still meeting the qualitative standards that we wanted to achieve. After the completion of the development of our project, we collected the results of MetricsReloaded in Table 5.5. The outliers, ConferenceController and StorageController, with high RFC values are not surprising since they are part of the intermediary layer and bundle functionalities in methods such as all database-related operations and managing the core classes of our data model. However, the low LCOM of these two classes indicates that they are cohesive and cannot be easily split into different classes. In general, we managed to stay below the limits for each metric, with notably low DIT, LCOM, and NOC, which indicates that our project consists of cohesive and reusable classes that are not too complex and, therefore, easy enough to maintain. In general, classes that have exceeded the aforementioned limits have these outliers are marked with $*$ besides the respective value and the class for easier visibility. However, the average of all classes do not exceed the boundaries. For this reason, we have reached our software evaluation goals.

[33] `https://plugins.jetbrains.com/plugin/93-metricsreloaded`, accessed: September 2022

| Class | CBO | DIT | LCOM | NOC | RFC | WMC |
|---|---|---|---|---|---|---|
| ConferencesApplication | 6 | 1 | 3 | 0 | 12 | 5 |
| ErrorFilter | 4 | 1 | 1 | 0 | 2 | 2 |
| data.metrics.APIMetric* | 6 | 2 | 2 | 1* | 16 | 7 |
| data.metrics.ApplicationType* | 9* | n/a | 3 | n/a | 12 | 8 |
| data.metrics.CalculatedMetric | n/a | n/a | n/a | n/a | 5 | n/a |
| data.metrics.ManualMetric | 2 | 2 | 4 | 0 | 7 | 6 |
| data.metrics.Metric* | 6 | 1 | 1 | 2* | 6 | 4 |
| data.metrics.zoom.AudioLatency | 7 | 3 | 5 | 0 | 17 | 6 |
| data.model.AdditionalMetric* | 11* | 1 | 4 | 0 | 19 | 16 |
| data.model.Conference* | 12* | 1 | 4 | 0 | 29 | 18 |
| data.model.ConferenceEdition* | 12* | 1 | 13 | 0 | 45 | 35* |
| data.model.IngestConfiguration* | 11* | 1 | 3 | 0 | 16 | 13 |
| data.model.dto.AdditionalMetricDTO | 7 | 1 | 4 | 0 | 45 | 24 |
| data.model.dto.ConferenceDTO | 5 | 1 | 5 | 0 | 27 | 14 |
| data.model.dto.ConferenceEditionDTO* | 4 | 1 | 2 | 0 | 78* | 38* |
| data.model.dto.ConferenceFrontendDTO | 3 | 1 | 1 | 0 | 46 | 29 |
| data.model.dto.DashboardingMetricDTO | 4 | 1 | 5 | 0 | 17 | 16 |
| data.model.dto.IngestConfigurationDTO | 6 | 1 | 2 | 0 | 35 | 18 |
| data.model.dto.SelectedMetricsDTO | 2 | 1 | 1 | 0 | 9 | 5 |
| data.util.Indices | 3 | n/a | 3 | n/a | 4 | 2 |
| server.controller.ConferenceController* | 18* | 1 | 2 | 0 | 102* | 30* |
| server.controller.Controller | n/a | n/a | n/a | n/a | 2 | n/a |
| server.controller.QoSMetricsController | 1 | 1 | 1 | 0 | 4 | 2 |
| server.controller.ServerController | 3 | 1 | 1 | 0 | 17 | 8 |
| controller.storage.ElasticOperation* | 3 | 1 | 1 | 2* | 2 | 2 |
| controller.storage.ElasticReadOperation | 8 | 2 | 1 | 0 | 36 | 12 |
| server.controller.storage.ElasticWriteOperation* | 9* | 2 | 1 | 0 | 54* | 35* |
| server.controller.storage.MapperService* | 13* | 1 | 2 | 0 | 29 | 11 |
| server.controller.storage.StorageController* | 19* | 1 | 2 | 0 | 115* | 81* |
| server.dashboarding.ChartType | 3 | n/a | 3 | n/a | 6 | 3 |
| server.dashboarding.DashboardingUtils | 7 | 1 | 1 | 0 | 42 | 21 |
| server.dashboarding.DashboardingUtils.<br>DashboardingConstants | 1 | 1 | 0 | 0 | 0 | 0 |
| server.dashboarding.Operations | 2 | n/a | 3 | n/a | 6 | 3 |
| server.rest.service.ConferenceService | n/a | n/a | n/a | n/a | 16 | n/a |
| server.rest.service.ConferenceServiceImpl* | 8 | 1 | 2 | 0 | 45 | 30* |
| server.rest.service.QoSMetricsService | n/a | n/a | n/a | n/a | 2 | n/a |
| server.rest.service.QoSMetricsServiceImpl* | 12* | 1 | 1 | 0 | 50* | 9 |
| **Average** | 6.14 | 0.95 | 2.35 | 0.14 | 26.35 | 13.86 |

**Table 5.5:** Software evaluation metrics calculated with MetricsReloaded

## 5.7 Installation of Reference Framework

This section handles how to proceed when using the tool itself.

### 5.7.1 Prerequisites

In order to install the reference framework, a few prerequisites must be met. First of all, the following have to be installed:

- Docker: Download and install the appropriate version of Docker for your operating system[34]
- Integrated Development Environment (IDE): An IDE of your choice has to be installed. If you want to use the MetricsReloaded plugin or if you want to follow our exact steps generally, then download and install the IntelliJ IDEA Community Edition[35].
- JDK: A JDK of your choice has to be installed. If you do not want to use Java 11, you need to adjust the Java version in the pom.xml of the reference framework. The framework is compatible with Java 11 and higher.
- Reference framework: Clone and import the project from this source[36]. Build the Maven project.

Optionally, the following plugins for IntelliJ are recommended:

- MetricsReloaded[37]: Install MetricsReloaded from Marketplace in IntelliJ.
- Docker[38]: Install MetricsReloaded from Marketplace in IntelliJ.

### 5.7.2 Docker Setup

Once the project prerequisites have been met, open up the project in the IDE. Additionally, open up Docker Desktop or the equivalent on your OS in the background. Navigate to the terminal tab in the IDE, and execute the command `docker compose up` directly into the terminal to set up the Docker containers with the configuration from `docker-compose.yml`. Note that executing this command is only necessary for a fresh install of the reference framework. To uninstall, type `docker compose down`.
After the Docker setup is finished, switch to Docker and check that "master project" and

[34] `https://docs.docker.com/get-docker/`, accessed: September 2022
[35] `https://www.jetbrains.com/idea/download/`, accessed: September 2022
[36] `https://github.com/lmhuber/masterproject`, accessed: September 2022
[37] `https://plugins.jetbrains.com/plugin/93-metricsreloaded`, accessed: September 2022
[38] `https://plugins.jetbrains.com/plugin/7724-docker`, accessed: September 2022

the associated containers have been successfully created, see Figure 5.15. If the containers are not already running after the setup, run master-project-kibana-1 and master project-elasticsearch-1 manually. These two containers must be running for the reference framework to function.

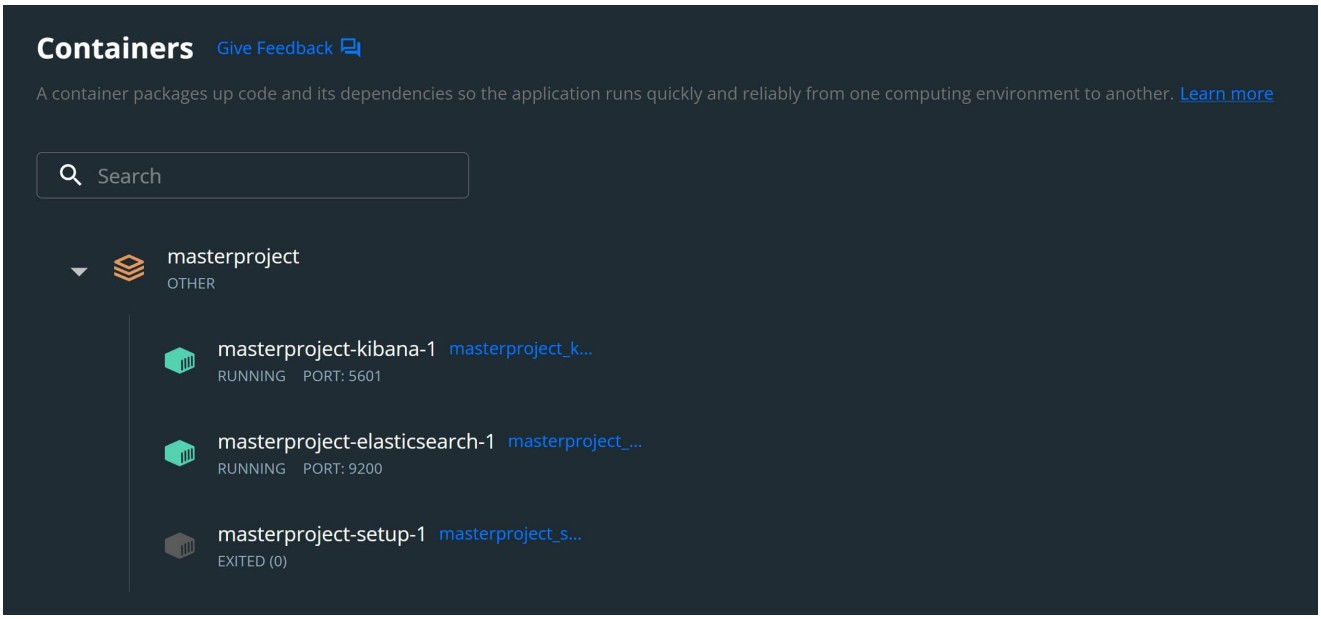



**Figure 5.15:** Overview of containers in Docker Desktop

### 5.7.3 Connecting to the Server

After a successful Docker setup, run the `ConferenceApplication` in the IDE. With both containers and the application running, open up a browser of your choice and enter `http://localhost:8080/conferences/list` onto the address bar. Spring Security will redirect you to the login page 5.17. You can log in with the credentials "user" and the random password Spring Security created on startup for demonstration purposes. You can find this password in the console; see Figure 5.16.

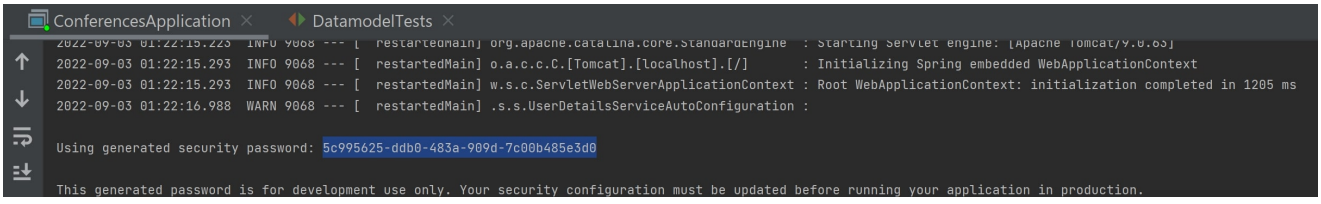



**Figure 5.16:** Autogenerated password by Spring Security

For logging into Kibana, enter `localhost:5601`, see Figure 5.18. The default credentials are "elastic" as the user name and "changeme" as the password. In order to change the password, navigate to `.env` and define a new one. Furthermore, a new API Key needs to be requested

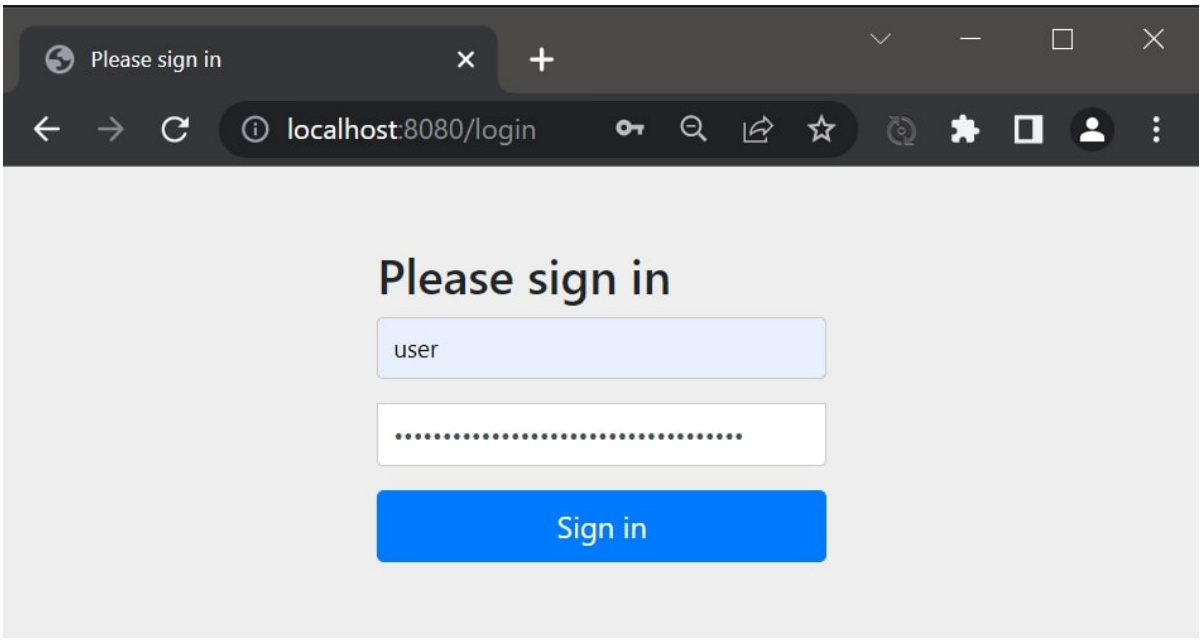



**Figure 5.17:** Log in to access the server

in Kibana. To achieve this, select "Stack Management," "API keys," and then "Create API key ."The newly created API key can be set in the `properties.configuration`.

After the login is successful, you are redirected to the Kibana Home page; see Figure. Navigate to Discover and select "conference" to view the current documents; see Figure 5.19.

To populate the database, change the following parameters in `ConferenceApplication`: `DEBUG` and `LOAD_SAMPLES` to true and run the `DatamodelTests`.

If necessary, Elasticsearch is directly reachable at `http://localhost:9200/`.

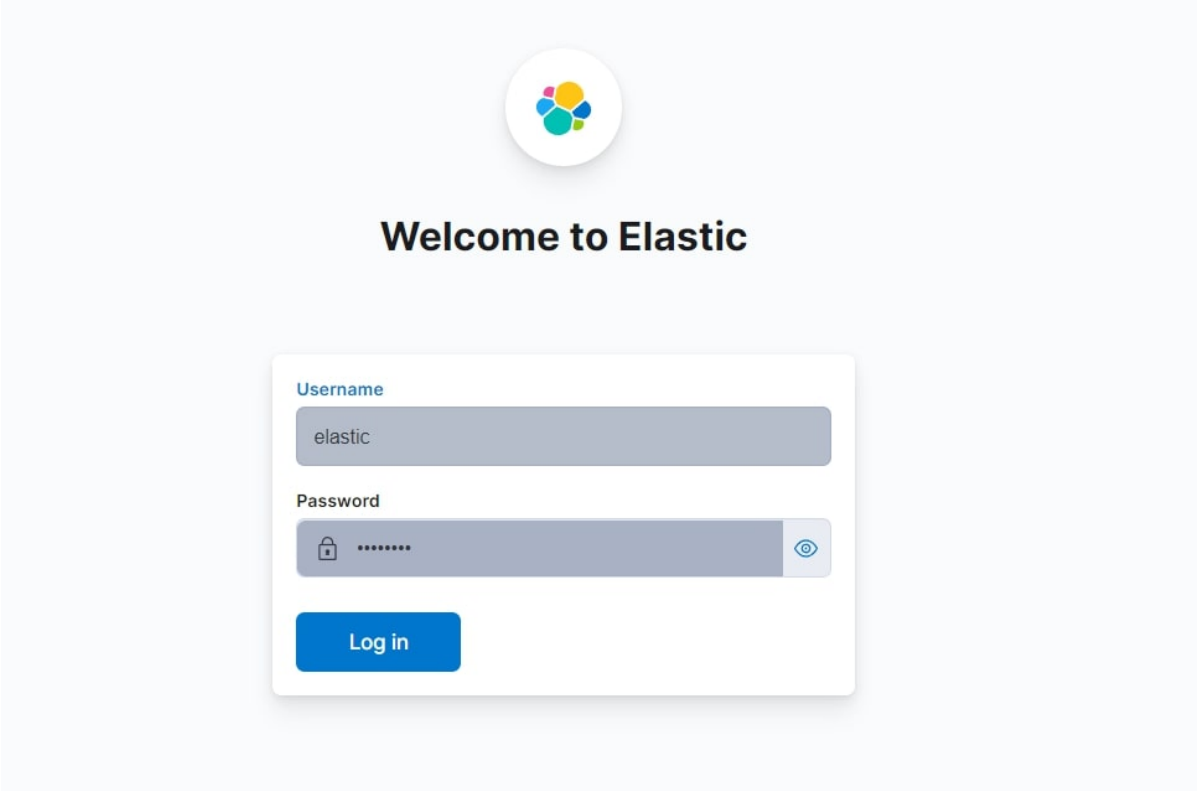



**Figure 5.18:** Log into Kibana

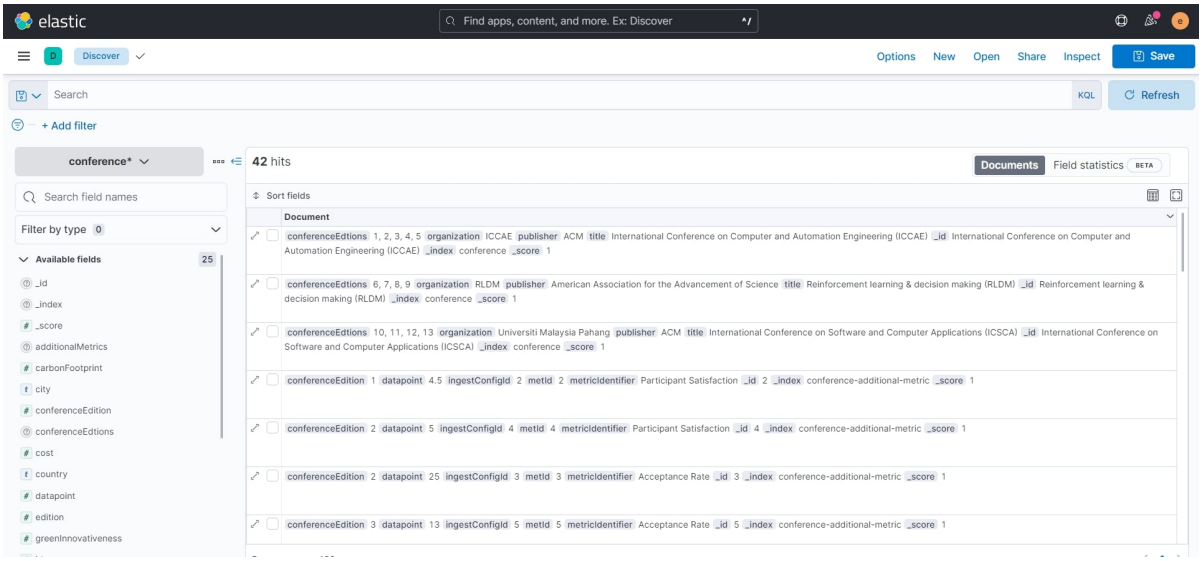



**Figure 5.19:** View Elasticsearch documents in Kibana

# 6 Results and Outlook

In this section, we provide a discussion of our results and give an outlook on directions future research should be headed, before we conclude the master thesis.

## 6.1 Results and Discussion

As Fulcher et al. [Ful+20] have noted, virtual conferences bring many benefits, such as a lower barrier to participation, but require careful planning. Lessons learned for them include that deficiency in holding virtual conferences has only come to light through retrospective analysis and not during the event itself [Ful+20]. In addition, Fulcher et al. [Ful+20] criticize that conference assessment, and evaluation is not well researched. This critique demonstrates the importance of delving deeper into this area of research. With our reference framework for evaluating virtual conferences, we are addressing precisely this gap.

Before heading into the detailed discussion, we want to provide an overview of the steps of the strategy we followed before discussing them later in detail. Building on an in-depth analysis of conferences mainly in the field of computer science, see Section 3, which also examined the differences between physical and virtual conferences, the basis for an evaluation framework for virtual conferences was established. In Section 4, we gradually answered the research questions we had formulated at the beginning, see Section 1, guided along three corresponding axes of a cube, see Figure 4.1. First, the anatomy of conferences with their core components and general requirements was assessed. Second, a set of evaluation criteria and their associated metrics was developed. Third, strategies and challenges in collecting data were identified and from these learnings, a refined metrics table and an evaluation criteria fingerprint for comparing virtual conferences was created. Ultimately, a reference framework for evaluating virtual conferences was designed and implemented.

Initially, we intended to develop automatically created evaluation reports for virtual conferences based on questionnaires, but abstained from that idea, once we discovered that Hohlfeld et al. [HGD21] had already pursued such an approach. Hohlfeld et al. [HGD21] took the following approach: Two questionnaires were designed for this survey. The first one was designed to determine the expectations and goals of the participants. The second questionnaire was designed to capture actual participation and participant experience in a data-driven

manner. The questionnaires can be analyzed with scripts so that, for example, Likert plots can be created using R. A key difference with our solution is that we can use the evaluation criteria to perform a much more comprehensive data analysis of a conference since reception, which was the focus area of the questionnaires, is only one aspect of several that should be considered during virtual conference evaluation. In addition, Hohlfeld et al. [HGD21] lack the possibility to compare and evaluate conferences for years, as we can with the dashboarding functionality of our reference framework as well as with the evaluation criteria fingerprint.

We will now discuss how we answered the research questions that we posed in Section 1. To iterate, we defined them as follows:

**RQ1** What is the anatomy of a conference, and how can it be visualized adequately with all parts and features?

**RQ2** What metrics can be used for evaluating virtual conferences?

**RQ3** How can we collect and compare the necessary data?

Concerning RQ1, we have analyzed the anatomy of a conference from the field of computer science and identified its core components, see Section 4.2. For this purpose, we designed use case diagrams for conferences from the perspective of different participants, e.g., attendants, organizers, and presenters, which visualize the processes. In terms of the core categorization, we have identified two main categories of actors, participant and organizational staff, as well as four distinct categories of action items, publication process, organizational affairs, conference tracks, and networking. Each of these categories groups together different actors or actions, and allows to split the conference into core components. With these core components, the characteristics and processes of a virtual conference can be easily understood. RQ2, which aimed at suitable metrics for the evaluation of virtual conferences, was answered by a comprehensive analysis of the factual basis and a subsequent definition of evaluation criteria, targeting key characteristics of a virtual conference, and their associated metrics, with which they can be measured and in turn evaluated, see Section 4.3. The following evaluation criteria were derived: Accessibility, usability, technical background, reception, economic aspects, ecological impact sustainability, privacy and security, and modality. Emphasis was placed on an overall view of all aspects of a virtual conference with additional focus on relevant topics like compliance with the GDPR and environmental impact and sustainability. The technical and reception qualities were also taken into account. Finally, a table of sample metrics was established for each evaluation criterion, see Table 4.1, which was further refined, see Table 4.4 after delving deeper into the topic of data collection in Section 4.4. These metrics cover essential aspects of a virtual conference and provide a holistic picture. However, not all sample metrics need to be used. This usage is only a recommendation and should be treated as a starting point for experts to select the right metrics for their individual virtual conference. Additional

custom metrics can be included at any time into the reference framework. As for RQ3, we answered this one in detail in Section 4.4. We have elaborated on all aspects of data collection, from definitions to concrete data collection procedures to the correct data set design, including challenges, such as the collection of green metrics, and risks, such as compliance to the GDPR and securing sensitive data. The collection of green metrics should be incorporated from the start into the conference policy, as data collection for these metrics is complex and relevant from the very beginning of organizing a virtual conference, since every meeting and used resource contributes to the carbon-dioxide footprint. Additionally, we discussed if and how data should be collected automatically. The answer is that automation should be definitely employed by researchers, but special consideration and care is necessary, when the automation process involves heterogeneous data sources, which need to be cleaned and made compatible properly. Even though automation is desirable, if data sources change regularly, the system can become error-prone. Therefore, manual inspections and testing processes are still necessary. One way to implement automation is to use CronJobs, which execute tasks at a repeating schedule.

Lastly, we have defined an evaluation criteria fingerprint that allows for comparison of virtual conferences based on key characteristics. We were inspired by U-Multirank[1] in the development of the fingerprint. The idea behind this fingerprint is to offer a visual aid to interpret all the key characteristics of a virtual conference, which we defined as the evaluation criteria we proposed, and to quickly be able to understand how well the conference performed in each area. With the help of the performance ranks that we defined, ranging from A (very good) to E (very weak), the performance can be interpreted and compared. For RQ3, we wanted to provide for a further option to visualize and compare the data of each edition of a virtual conference over the years. In order to achieve this, we decided to include dashboarding into our reference framework, which allows organizers to not only see the data for each selected metric in the form of diagrams, but also to see historic data in the form of a time series. Therefore, organizers can learn from past editions of a virtual conference and improve it iteration by iteration. After answering our research questions, we devoted Section 5 to the practical case study with the implementation of the reference framework. Regarding the requirements analysis, see Section 5. In comparison with other evaluation solutions, see Section 5.1, we decided to focus on dashboards, which represent interactive reports for conferences, see Section 5.4.4. In this way, the reference framework is even more flexible since reports do not have to be generated in a cumbersome manner manually. However, desired metrics or widgets, e.g., displaying histograms, can be added or removed simply by clicking. In addition, the visual presentation of the information is thus better provided. Moreover, reports can be generated from the dashboards using Kibana, provided that a Kibana subscription is available[2].

[1] `https://www.umultirank.org/`, accessed: September 2022

[2] `https://www.elastic.co/guide/en/kibana/current/reporting-getting-started.html`, accessed: September 2022

Furthermore, we want to clearly distinguish our reference framework from the tools developed by Bartindale et al. [Bar+21] as well as commercial software solutions such as Flowmon [Cor20], Bluejeans by Verizon [Ver22], and Read.ai [AI22]. Bartindale et al. [Bar+21] developed the ZoomSense tool, a scalable infrastructure that adds some functionality to Zoom. In particular, it monitors technical and reception metrics. Consent handling was particularly influential, as it allowed users to opt-out of data collection at any time. Similarly, ZoomSense's system architecture inspired us to consider a microservices architecture. However, after careful consideration, we chose a small monolith architecture, which was a better fit for the requirements of our tool, see Section 5.6.1. Even though ZoomSense is an exciting tool, the use case is also limited to conferences that rely on Zoom. A significant difference to our reference framework is that we also allow custom metrics, and the application can be tailored to various conferences. So not only is there no dependency on specific tools, but the focus can also be arbitrarily placed on other metrics, not just technical and reception metrics. Therefore, our reference framework creates an overall view of a conference. Flowmon [Cor20] is again geared towards analyzing the technical and interaction quality of the participants in a meeting, e.g., Zoom or Webex meetings. BlueJeans from Verizon [Ver22] is a video conferencing platform with dashboarding functionalities. Apart from technical and interaction metrics, some economic data, such as ROI, can also be collected. Read.ai [AI22] is designed for the quantitative evaluation of a video conference, where, e.g., audio quality and viewer engagement are decisive for a meeting score. The problem with all three commercial solutions is that they focus only on the technical and interaction quality of the video conference. They are not tools specifically designed to evaluate academic conferences. Furthermore, there is no possibility of defining custom metrics. All other evaluation solutions also lack metrics that assess environmental impact and sustainability. The areas of accessibility and usability, as well as privacy and security, are also not addressed.

Based on the requirements defined in Section 5, in consultation with our supervisor, we will now examine aspects of the reference framework in more detail. The reference framework has a UI built using Thymeleaf and Bootstrap, which allows listing and searching for conferences. Implementing the ELK stack and Docker setup makes it possible to persist data over the years. In addition, editions can be created for conferences to store data for each event. It is possible to search for conferences in the full-text search or filter them by criteria using the advanced search. In addition, they can be sorted in descending or ascending order according to the respective values. Fields can also be hidden to limit the search to the relevant fields and thus facilitate comparison between conferences. Another requirement was to show a conference's history and the evolution of metrics across different editions. For this purpose, only the dashboard for the respective conference has to be called up. It is possible to select in advance which additional metrics will be displayed. With custom metrics and custom tiles, the dashboard can be customized to show precisely the visualizations of data that the user is interested in. Users can also manually create conferences and add additional metrics from

any number-based data points and with any labels. A cron job for reading metrics from the mocked Zoom API has also been implemented as a proof of concept. This test allows for automatic and periodic metrics reading as long as the server and the mocked API are running. Every day at 00:00, the metrics are read and automatically stored in the ElasticSearch repository.

The original plan was to read metrics directly via the Zoom API[3]. However, it turned out that the relevant metrics on technical and interaction aspects of the meeting are only available with the most expensive business plans. For this reason, it was decided to mock the Zoom API for our unit tests, based on the official specification[4]. In this way, the official endpoints can be used, and, in the case where an appropriate Zoom license has been purchased, only the URL needs to be exchanged.

The reference framework can capture and observe all the conference details over the years. Organizers have insight into the development of the conference and can determine whether the measures taken have an effect over time or which adjustments need to be made. The dashboard's visualizations help to interpret these developments and, if necessary, limit them precisely to the relevant years or key points by making a selection. Of course, the metrics collected alone are not enough to determine what actions need to be taken. This evaluation still requires the expert knowledge of the organizers of a conference. Nevertheless, they certainly help to ensure that a data-based decision is made. By meta-studying the literature and analyzing our factual base, we captured those evaluation criteria that cover all the essential aspects of virtual conferences. In addition, we identified metrics that can be used to observe those aspects that are particularly relevant for the success of a virtual conference. Furthermore, using the reference framework can help examine research in comparing different conference formats and their effectiveness and sustainability. The tool can also monitor the carbon footprint of conferences and explore which measures are proving most effective on metrics.

## 6.2 Future Work

Generally speaking, there is still much research on virtual conferencing, as stated by Fulcher et al. [Ful+20]. Some particular areas of development have emerged through the analysis, which should be the focus of future research. One topic that concerns not only virtual conferences but conferences in general, is the reduction of the carbon footprint and sustainability. Especially scientists who have a pioneering role in society should actively contribute to reducing the considerable carbon footprint of conferences. Our analysis has shown that physical conferences

[3] `https://marketplace.zoom.us/docs/api-reference/zoom-api/methods/`, accessed: September 2022
[4] `https://marketplace.zoom.us/docs/api-reference/zoom-api/methods/`, accessed: September 2022

offer several advantages that virtual conferences cannot easily replace. In our view, physical conferences will not be completely replaced by virtual conferences but supplemented. In particular, there is also a need for comparisons between physical and virtual conferences or hybrid forms. Specifically, the social aspect requires further development and new solutions, be it through the simulation of eye contact via 3D videos [He+21; TWL20] or through tools such as Minglr [SRM21], which simulate spontaneous meetings with a wide variety of conference participants. Another important aspect is that research is needed into virtual meeting formats that can be established as standards. Overall, the COVID-19 pandemic has increased the use of virtual conferences and triggered their further development. However, research is still lacking compared to physical conferences, which have been the dominant format in the past.

Concerning the further development of our reference framework, the area of automation should be advanced so that more data can be collected and read automatically. For this master thesis, we have integrated automation only as a proof of concept. However, we are convinced that it is the right way to develop the tool further since automation reduces the workload for researchers. Another critical step is to deploy the framework in the cloud or on a server so that the cron job can run automatically and independently from the device. Another aspect that should not be neglected is security. With Spring Security, important guidelines have already been established. Nevertheless, the endpoints should still be restricted by access protocols and user access by authentication measures. Additionally, the framework could be improved by integrating the functionality to add custom ranges for metrics.
In terms of further development of the evaluation criteria fingerprint, we could see it as an expansion to the reference framework, so that metrics could be selected directly from the available data in the framework and generate a fingerprint automatically. Moreover, a way to directly display and compare fingerprints of different conference editions would be desirable. From a metrics perspective, as Zhang et al. [Zha+21] points out, measuring interaction dynamics using an interactive long short-term memory neural network (LSTM) with conversational sentiment analysis would an interesting factor to consider.

## 6.3 Conclusion

In summary, we surveyed the state of the art of virtual conferences in Computer Science using a meta-study, see Section 3. After an in-depth analysis in Section 4, we elicited the requirements for the reference framework based on three dimensions. The first dimension, see Section 4.2, covered the anatomy of a conference with the identification of core components and core processes. In the second dimension, see Section 4.3, evaluation criteria were determined by extracting the meta-study results and defining them. Multiple metrics were defined for each evaluation criterion and ultimately captured in an overview table. In Section 4.4, we

explored the third dimension, which concerns data collection for metrics. For this purpose, we examined data collection techniques that can be considered for our reference framework. We also addressed specific issues, such as the challenges of data collection concerning green metrics or the risks of data collection related to the GDPR. Then, based on the results, we established a refined set of sample metrics that can be used as a starting point for selecting appropriate metrics for a virtual conference. Additionally, we put forward the concept for an evaluation criteria fingerprint, see Section 4.4.11, which can be utilized to see the characteristics of a virtual conference at a glance and visually interpret how good the performance was. The fingerprint can be also used as a visual aid to compare editions of virtual conferences. In comparison with other evaluation frameworks, see Section 5.1 and 6.1, we have developed a framework that can be used generally for a wide variety of conferences, and we have defined a more comprehensive range of evaluation criteria and metrics. Our framework is also extensible with custom metrics and provides an interactive dashboard with the ability to get an overview of historical conference data. In addition, we included automated data collection of metrics from a mocked Zoom API as a proof of concept.

Our contribution, therefore, is that, on the one hand, we have advanced research on virtual conferences and how they can be assessed and evaluated. On the other hand, we have developed a reference framework that can serve as a basis for the practical evaluation of virtual conferences. The reference framework lays the groundwork for future studies to assess the performance of virtual conferences. It is a flexibly extensible tool that can be used for various conferences and extended by any number-based metrics. This fact, as well as the expandability from an implementation point of view, is shown in Section 5.6.

# Bibliography

[Ach+21] Titipat Achakulvisut et al. "Towards Democratizing and Automating Online Conferences: Lessons from the Neuromatch Conferences". In: *Trends in Cognitive Sciences*. Vol. 25. 4. Neuromatch Conferences. Elsevier BV, Apr. 2021, pp. 265–268. DOI: 10.1016/j.tics.2021.01.007. URL: https://doi.org/10.1016/j.tics.2021.01.007.

[AI22] Read AI. *Read Dashboard*. 2022. URL: https://www.read.ai/.

[Alo+20] Lucia Alonso-Virgos, Jordán Pascual Espada, Oscar Sanjuán Martínez, and Rubén González Crespo. "Compliance and application tests of usability guidelines about giving information quickly and comprehensibly". In: Springer, 2020, pp. 1–21. DOI: https://doi.org/10.1007/s40747-020-00198-5.

[Alw20] Alaa A. Alwabel. "Privacy Issues in Big Data from Collection to Use". In: *Big Data and Security*. Ed. by Yuan Tian, Tinghuai Ma, and Muhammad Khurram Khan. Singapore: Springer Singapore, 2020, pp. 382–391. ISBN: 978-981-15-7530-3.

[Arc+19] Mandy M. Archibald, Rachel C. Ambagtsheer, Mavourneen G. Casey, and Michael Lawless. "Using Zoom Videoconferencing for Qualitative Data Collection: Perceptions and Experiences of Researchers and Participants". In: *International Journal of Qualitative Methods*. Vol. 18. Sage Pub, 2019, p. 1609406919874596. DOI: 10.1177/1609406919874596. eprint: https://doi.org/10.1177/1609406919874596. URL: https://doi.org/10.1177/1609406919874596.

[Bar+21] Tom Bartindale, Peter Chen, Harrison Marshall, Stanislav Pozdniakov, and Dan Richardson. "ZoomSense: A Scalable Infrastructure for Augmenting Zoom". In: *Proceedings of the 29th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 3771–3774. ISBN: 9781450386517. URL: https://doi.org/10.1145/3474085.3478332.

[BBM96] V.R. Basili, L.C. Briand, and W.L. Melo. "A validation of object-oriented design metrics as quality indicators". In: *IEEE Transactions on Software Engineering* 22.10 (1996), pp. 751–761. DOI: 10.1109/32.544352.

[Bie+21] Richard Bieringa et al. "An Empirical Evaluation of the Performance of Video Conferencing Systems". In: *Companion of the ACM/SPEC International Conference on Performance Engineering*. ICPE '21. Virtual Event, France: Association for Computing Machinery, 2021, pp. 65–71. ISBN: 9781450383318. DOI: 10.1145/3447545.3451186. URL: https://doi.org/10.1145/3447545.3451186.

[Bil+20] Angelos Bilas et al. *The EuroSys 2020 Online Conference: Experience and lessons learned*. EuroSys Conference. 2020. arXiv: 2006.11068 [cs.CY].

[BK20] Beatrix Busse and Ingo Kleiber. "Realizing an online conference: Organization, management, tools, communication, and co-creation". In: vol. 25. 3. International Computer Archive of Modern and Medieval English (ICAME) Conference. John Benjamins, 2020, pp. 322–346. DOI: https://doi.org/10.1075/ijcl.00028.bus. URL: https://www.jbe-platform.com/content/journals/10.1075/ijcl.00028.bus.

[Blo78] Arthur Bloch. *Murphy's Law and Other Reasons Why Things Go Wrong*. Price / Stern / Sloan Publishers, 1978. ISBN: 9780843104288. URL: https://www.penguinrandomhouse.ca/books/291795/murphys-law-the-26th-anniversary-edition-by-arthur-bloch/9781101663752.

[Bon+21] Angela Bonifati et al. "Holding a Conference Online and Live Due to Covid-19: Experiences and Lessons Learned from EDBT / ICDT 2020". In: *SIGMOD Rec.* Vol. 49. 4. ICDT / EBT. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 28–32. DOI: 10.1145/3456859.3456866. URL: https://doi.org/10.1145/3456859.3456866.

[Bon+22] Philippe Bonnet, Xin Luna Dong, Felix Naumann, and Pinar Tözün. "VLDB 2021: Designing a Hybrid Conference". In: *SIGMOD Rec.* Vol. 50. 4. International Conference on Very Large Databases (VLDB). New York, NY, USA: Association for Computing Machinery, Jan. 2022, pp. 50–53. DOI: 10.1145/3516431.3516447. URL: https://doi.org/10.1145/3516431.3516447.

[CK94] S.R. Chidamber and C.F. Kemerer. "A metrics suite for object oriented design". In: *IEEE Transactions on Software Engineering* 20.6 (1994), pp. 476–493. DOI: 10.1109/32.295895.

[CMR21] Pasquale Cantiello, Michele Mastroianni, and Massimiliano Rak. "A Conceptual Model for the General Data Protection Regulation". In: *Computational Science and Its Applications – ICCSA 2021*. Ed. by Osvaldo Gervasi et al. Cham: Springer International Publishing, 2021, pp. 60–77. ISBN: 978-3-030-87010-2.

[Cor20] Progress Software Corporation. *Tracking the Performance of Online Meeting Tools*. 2020. URL: https://www.flowmon.com/en/blog/tracking-the-performance-of-online-meeting-tools.

[DD18] Kevin Doherty and Gavin Doherty. "Engagement in HCI: Conception, Theory and Measurement". In: *ACM Comput. Surv.* Vol. 51. 5. New York, NY, USA: Association for Computing Machinery, Nov. 2018. DOI: 10.1145/3234149. URL: https://doi.org/10.1145/3234149.

[De +19] Berardina De Carolis, Francesca D'Errico, Nicola Macchiarulo, and Giuseppe Palestra. ""Engaged Faces": Measuring and Monitoring Student Engagement from Face and Gaze Behavior". In: *IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume*. WI '19 Companion. Thessaloniki, Greece: Association for Computing Machinery, 2019, pp. 80–85. ISBN: 9781450369886. DOI: 10.1145/3358695.3361748. URL: https://doi.org/10.1145/3358695.3361748.

[DF08] Jackie L. Dobrovolny and Stephanie Christine G. Fuentes. "Quantitative versus qualitative evaluation: A tool to decide which to use". In: vol. 47. 4. 2008, pp. 7–14. DOI: https://doi.org/10.1002/pfi.197. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pfi.197. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pfi.197.

[Dia+22] Vasiliki Diamantopoulou, Costas Lambrinoudakis, Jennifer King, and Stefanos Gritzalis. "EU GDPR: Toward a Regulatory Initiative for Deploying a Private Digital Era". In: *Modern Socio-Technical Perspectives on Privacy*. Ed. by Bart P. Knijnenburg et al. Cham: Springer International Publishing, 2022, pp. 427–448. ISBN: 978-3-030-82786-1. DOI: 10.1007/978-3-030-82786-1_18. URL: https://doi.org/10.1007/978-3-030-82786-1_18.

[DTK20] Vasiliki Diamantopoulou, Aggeliki Tsohou, and Maria Karyda. "From ISO/IEC 27002:2013 Information Security Controls to Personal Data Protection Controls: Guidelines for GDPR Compliance". In: *Computer Security*. Ed. by Sokratis Katsikas et al. Cham: Springer International Publishing, 2020, pp. 238–257. ISBN: 978-3-030-42048-2.

[DZM20] Mario Diethart, Anne Zimmermann, and Ingrid Mulà. "Guidelines for virtual conferencing–inspired by the COPERNICUS Alliance Online Conference 2019". In: COPERNICUS Alliance Online Conference. COPERNICUS Alliance, 2020. URL: https://boris.unibe.ch/139254/1/Zimmermann_2020_Guidelines%20for%20Virtual%20Conferencing.pdf.

[Ede+17] Johan Edelheim, Kimberly Thomas-Francois, Kajsa Åberg, and Giang Phi. "What do conferences do? What is academics' intangible return on investment (ROI) from attending an academic tourism conference?" In: *Journal of Teaching in Travel & Tourism*. Vol. 18. TEFI 9 - Celebrating the Disruptive Power of Caring in Tourism Education. Taylor & Francis, Dec. 2017, pp. 1–14. DOI: 10.1080/15313220.2017.1407517.

[Ela22] ElasticSearch B.V. *ElasticSearch Website*. Online. https://www.elastic.co (accessed: September 2022). 2022.

[Ell17] Justin Ellingwood. *An Introduction to Metrics, Monitoring, and Alerting*. 2017. URL: https://www.digitalocean.com/community/tutorials/an-introduction-to-metrics-monitoring-and-alerting.

[Fab21] Grant Faber. "A framework to estimate emissions from virtual conferences". In: *International Journal of Environmental Studies*. Vol. 78. 4. Routledge, 2021, pp. 608–623. DOI: 10.1080/00207233.2020.1864190. eprint: https://doi.org/10.1080/00207233.2020.1864190. URL: https://doi.org/10.1080/00207233.2020.1864190.

[For+21] Joël Foramitti, Stefan Drews, Franziska Klein, and Théo Konc. "The virtues of virtual conferences". In: *Journal of Cleaner Production*. Vol. 294. ICTA Virtual Conference. Elsevier, 2021, p. 126287. DOI: https://doi.org/10.1016/j.jclepro.2021.126287. URL: https://www.sciencedirect.com/science/article/pii/S0959652621005072.

[Fox21] Harry J. Foxwell. *Creating Good Data. A Guide to Dataset Structure and Data Representation*. Apress Berkeley, CA, 2021. ISBN: 978-1-4842-6103-3. URL: https://doi.org/10.1007/978-1-4842-6103-3.

[Ful+20] Michael R. Fulcher et al. "Broadening Participation in Scientific Conferences during the Era of Social Distancing". In: *Trends in Microbiology*. Vol. 28. 12. Natural Science Foundation - Deciphering the Microbiome. Elsevier BV, Dec. 2020, pp. 949–952. DOI: 10.1016/j.tim.2020.08.004. URL: https://doi.org/10.1016/j.tim.2020.08.004.

[Gar20] Moisés Macero García. *Learn Microservices with Spring Boot. A Practical Approach to RESTful Services Using an Event-Driven Architecture, Cloud-Native Patterns, and Containerization*. 2nd ed. Apress Berkeley, CA, 2020. ISBN: 978-1-4842-6130-9. URL: https://doi.org/10.1007/978-1-4842-6131-6.

[GCR16] P. Gómez, R. Casallas, and C. Roncancio. "Data schema does matter, even in NoSQL systems!" In: *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*. 2016, pp. 1–6. DOI: 10.1109/RCIS.2016.7549340.

[GML20] Zafeiroula Georgiopoulou, Eleni-Laskarina Makri, and Costas Lambrinoudakis. "GDPR Compliance: Proposed Technical and Organizational Measures for Cloud Providers". In: *Computer Security*. Ed. by Sokratis Katsikas et al. Cham: Springer International Publishing, 2020, pp. 181–194. ISBN: 978-3-030-42048-2.

[Gra22] Grafana Labs. *Grafana Website*. Online. https://grafana.com (accessed: September 2022). 2022.

[Has+21] Muhammad Hasnain, Muhammad Fermi Pasha, Imran Ghani, and Seung Ryul Jeong. "Simulated dataset collection method of dynamic quality of services (QoS) metrics". In: 2021, pp. 1–7. URL: https://link.springer.com/article/10.1007/s41870-021-00647-5.

[He+21] Zhenyi He, Keru Wang, Brandon Yushan Feng, Ruofei Du, and Ken Perlin. "GazeChat: Enhancing Virtual Conferences with Gaze-Aware 3D Photos". In: *The 34th Annual ACM Symposium on User Interface Software and Technology*. UIST '21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 769–782. ISBN: 9781450386357. DOI: 10.1145/3472749.3474785. URL: https://doi.org/10.1145/3472749.3474785.

[Hei+20] Vanessa Heitplatz, Cosima Nellen, Lena Sube, and Christian Bühler. "Implementing New Technological Devices in Social Services: Introducing the miTAS Project". In: vol. 1. ICCHP, Sept. 2020, pp. 109–118. URL: https://re.public.polimi.it/retrieve/handle/11311/1151879/556591/ED_1_ICCHP_Forum.pdf#page=109.

[Hen] Shawn Lawton Henry. *Web Accessibility Intro*. https://www.w3.org/WAI/fundamentals/accessibility-intro/ (accessed: September 2022).

[HGD21] Oliver Hohlfeld, Dennis Guse, and Katrien De Moor. "A Questionnaire to Assess Virtual Conference Participation Experience". In: *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2021, pp. 197–200. DOI: 10.1109/QoMEX51781.2021.9465406.

[HH21] Raiful Hasan and Ragib Hasan. "Towards a Threat Model and Security Analysis of Video Conferencing Systems". In: *2021 IEEE 18th Annual Consumer Communications Networking Conference (CCNC)*. IEEE Annual Consumer Communications Networking Conference. 2021, pp. 1–4. DOI: 10.1109/CCNC49032.2021.9369505.

[Hou+04] Robert J House, Paul J Hanges, Mansour Javidan, Peter W Dorfman, and Vipin Gupta. *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications, 2004. ISBN: 1452208123.

[HWK22] William Hurst, Adam Withington, and Hoshang Kolivand. "Virtual conference design: features and obstacles". In: *Multimedia Tools and Applications*. Springer, Mar. 2022. DOI: 10.1007/s11042-022-12402-4. URL: https://link.springer.com/article/10.1007/s11042-022-12402-4.

[HXR21] Maria Haji-Georgi, Xinyun Xu, and Oxana Rosca. "Academic conferencing in 2020: A virtual conference model". In: *Human Behavior and Emerging Technologies*. Vol. 3. 1. Wiley, 2021, pp. 176–184. DOI: https://doi.org/10.1002/hbe2.235. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbe2.235. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbe2.235.

[Ill21] Anusha Illukkumbura. *Introduction to Time Series Analysis*. Independently published, 2021. ISBN: 9798769616464. URL: http://www.anushabooks.com/product/introduction-to-time-series-analysis/.

[Jør18] Magne Jørgensen. "Do Agile Methods Work for Large Software Projects?" In: *Agile Processes in Software Engineering and Extreme Programming*. Ed. by Juan Garbajosa, Xiaofeng Wang, and Ademar Aguiar. Cham: Springer International Publishing, 2018, pp. 179–190. ISBN: 978-3-319-91602-6.

[Jos+15] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. "Likert scale: Explored and explained". In: vol. 7. 4. SCIENCEDOMAIN International, 2015, p. 396. URL: https://eclass.aspete.gr/modules/document/file.php/EPPAIK269/5a7cc366dd963113c6923ac4a73c3286ab22.pdf.

[Käv+22] Joakim Kävrestad, Allex Hagberg, Robert Roos, Jana Rambusch, and Marcus Nohlberg. "Usable Privacy and Security from the Perspective of Cognitive Abilities". In: *Privacy and Identity Management. Between Data Protection and Security*. Ed. by Michael Friedewald, Stephan Krenn, Ina Schiering, and Stefan Schiffner. Cham: Springer International Publishing, 2022, pp. 105–121. ISBN: 978-3-030-99100-5.

[Kho+20] Selina Khoirom, Moirangthem Sonia, Borishphia Laikhuram, Jaeson Laishram, and Tekcham Davidson Singh. "Comparative analysis of Python and Java for beginners". In: vol. 7. 8. 2020, pp. 4384–4407. URL: https://d1wqtxts1xzle7.cloudfront.net/64732739/IRJET_V7I8755-with-cover-page-v2.pdf.

[Khu12] Amandeep Khurana. "Introduction to HBase schema design". In: *White Paper, Cloudera*. 2012.

[Kin04] Zella King. "Career self-management: Its nature, causes and consequences". In: vol. 65. 1. Elsevier, 2004, pp. 112–133. DOI: https://doi.org/10.1016/S0001-8791(03)00052-6. URL: https://www.sciencedirect.com/science/article/pii/S0001879103000526.

[Kle17] Martin Kleppmann. *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems*. " O'Reilly Media, Inc.", 2017.

[KLK20] Samuel Kounev, Klaus-Dieter Lange, and Jóakim von Kistowski. *Systems Benchmarking. For Scientists and Engineers*. Springer Cham, 2020. ISBN: 978-3-030-41704-8. URL: https://doi.org/10.1007/978-3-030-41705-5.

[KW19] Pwint Phyu Khine and Zhaoshun Wang. "A Review of Polyglot Persistence in the Big Data World". In: vol. 10. 4. 2019. DOI: 10.3390/info10040141. URL: https://www.mdpi.com/2078-2489/10/4/141.

[Lah+21] Saadi Lahlou et al. "Are We 'Beyond Being There' yet? Towards Better Interweaving Epistemic and Social Aspects of Virtual Reality Conferencing". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Stanford media-X Global Innovation Leadership Program (GILP). New York, NY, USA: Association for Computing Machinery, 2021. ISBN: 9781450380959. URL: https://doi.org/10.1145/3411763.3451579.

[LC20] Dudley W. Lamming and Christy S. Carter. "Maintaining a scientific community while social distancing". In: *Translational Medicine of Aging*. Vol. 4. Aging Science Talks: Science for the community. 2020, pp. 55–59. DOI: https://doi.org/10.1016/j.tma.2020.05.002. URL: https://www.sciencedirect.com/science/article/pii/S2468501120300067.

[LMO20] Duc Anh Le, Blair MacIntyre, and Jessica Outlaw. "Enhancing the Experience of Virtual Conferences in Social Virtual Environments". In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2020, pp. 485–494. DOI: 10.1109/VRW50115.2020.00101.

[Mac+12] Bjoern-Elmar Macek, Christoph Scholz, Martin Atzmueller, and Gerd Stumme. "Anatomy of a Conference". In: *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*. HT '12. Hypertext Conference. Milwaukee, Wisconsin, USA: Association for Computing Machinery, 2012, pp. 245–254. ISBN: 9781450313353. DOI: 10.1145/2309996.2310038. URL: https://doi.org/10.1145/2309996.2310038.

[Mac12] Linda A Macaulay. *Requirements engineering*. Springer Science & Business Media, 2012. DOI: 10.1007/978-1-4471-1005-7.

[MB16] N Menold and K Bogner. "Design of rating scales in questionnaires". In: GESIS–Leibniz Institute for the Social Sciences Mannheim, Germany, 2016, p. 4. URL: https://www.gesis.org/fileadmin/upload/SDMwiki/MenoldBogner_Design_of_Rating_Scales_in_Questionnaires.pdf.

[Mey+21] Michael F. Meyer et al. "Virtual Growing Pains: Initial Lessons Learned from Organizing Virtual Workshops, Summits, Conferences, and Networking Events during a Global Pandemic". In: *Limnology and Oceanography Bulletin*. Vol. 30. 1. GLEON, NALMS, PPNW, EFI, KGML, DSOS. Wiley, 2021, pp. 1–11. DOI: https://doi.org/10.1002/lob.10431. eprint: https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.1002/lob.10431. URL: https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.1002/lob.10431.

[MGL20] Eleni-Laskarina Makri, Zafeiroula Georgiopoulou, and Costas Lambrinoudakis. "A Proposed Privacy Impact Assessment Method Using Metrics Based on Organizational Characteristics". In: *Computer Security*. Ed. by Sokratis Katsikas et

al. Cham: Springer International Publishing, 2020, pp. 122–139. ISBN: 978-3-030-42048-2.

[Mis+20] Chris Misa et al. "Lessons Learned Organizing the PAM 2020 Virtual Conference". In: *SIGCOMM Comput. Commun. Rev.* Vol. 50. 3. Passive and Active Measurement Conference. New York, NY, USA: Association for Computing Machinery, July 2020, pp. 46–54. DOI: 10.1145/3411740.3411747. URL: https://doi.org/10.1145/3411740.3411747.

[MK13] Innocent Mapanga and Prudence Kadebu. "Database management systems: A nosql analysis". In: vol. 1. 7. Engineering Research Publication, 2013, p. 265849. URL: https://www.researchgate.net/profile/Innocent-Mapanga/publication/258328266_Database_Management_Systems_A_NoSQL_Analysis/links/00b49527d04f2f19ed000000/Database-Management-Systems-A-NoSQL-Analysis.pdf.

[MK14] Dan McCreary and Ann Kelly. "Making sense of NoSQL". In: 2014, pp. 19–20. URL: http://www.dreamtechpress.com/wp-content/uploads/2019/02/toc_978-93-5119-202-2.pdf.

[MNM20] Arthur-Jozsef Molnar, Alexandra NeamŢu, and Simona Motogna. "Evaluation of Software Product Quality Metrics". In: *Evaluation of Novel Approaches to Software Engineering*. Ed. by Ernesto Damiani, George Spanoudakis, and Leszek A. Maciaszek. Cham: Springer International Publishing, 2020, pp. 163–187. ISBN: 978-3-030-40223-5.

[MNM21] Andrew N. Mason, John Narcum, and Kevin Mason. "Social media marketing gains importance after Covid-19". In: ed. by Usama Awan. Vol. 8. 1. Cogent OA, 2021, p. 1870797. DOI: 10.1080/23311975.2020.1870797. eprint: https://doi.org/10.1080/23311975.2020.1870797. URL: https://doi.org/10.1080/23311975.2020.1870797.

[Mon22] MongoDB, Inc. *MongoDB Website*. Online. https://www.mongodb.com/ (accessed: September 2022). 2022.

[NAT21] Abdallah Namoun, Ahmed Alrehaili, and Ali Tufail. "A Review of Automated Website Usability Evaluation Tools: Research Issues and Challenges". In: *Design, User Experience, and Usability: UX Research and Design*. Ed. by Marcelo M. Soares, Elizabeth Rosenzweig, and Aaron Marcus. Cham: Springer International Publishing, 2021, pp. 292–311. ISBN: 978-3-030-78221-4.

[Neu+20] Sabrina Neugebauer, Maren Bolz, Rose Mankaa, and Marzia Traverso. "How sustainable are sustainability conferences? – Comprehensive Life Cycle Assessment of an international conference series in Europe". In: *Journal of Cleaner Production*. Vol. 242. 3-day international academic conference on sustainability topics. Elsevier, 2020, p. 118516. DOI: https://doi.org/10.1016/j.jclepro.2019.118516.

URL: https://www.sciencedirect.com/science/article/pii/S0959652619333864.

[OA20] Peterson K Ozili and Thankom Arun. "Spillover of COVID-19: impact on the Global Economy". In: SSRN, 2020. URL: https://dx.doi.org/10.2139/ssrn.3562570.

[OBT21] Egil Ovrelid, Bendik Bygstad, and Gard Thomassen. "TSD: A Research Platform for Sensitive Data". In: *Procedia Computer Science*. Vol. 181. CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020. Elsevier, 2021, pp. 127–134. DOI: https://doi.org/10.1016/j.procs.2021.01.112. URL: https://www.sciencedirect.com/science/article/pii/S1877050921001496.

[Ous18] John Ousterhout. *A Philosophy of Software Design*. Palo Alto, CA: Yaknyam, 2018. ISBN: 978-1-73210-220-0.

[Pal21] Chris Palmer. "Are Virtual Conferences Here to Stay?" In: *Engineering*. Elsevier, 2021. DOI: https://doi.org/10.1016/j.eng.2021.01.005. URL: https://www.sciencedirect.com/science/article/pii/S2095809921000126.

[Par+21] Richard Parncutt, PerMagnus Lindborg, Nils Meyer-Kahlen, and Renee Timmers. "The Multi-hub Academic Conference: Global, Inclusive, Culturally Diverse, Creative, Sustainable". In: vol. 6. 2021. DOI: 10.3389/frma.2021.699782. URL: https://www.frontiersin.org/article/10.3389/frma.2021.699782.

[Pat+20] Gourab K Patro, Abhijnan Chakraborty, Niloy Ganguly, and Krishna P. Gummadi. *On Fair Virtual Conference Scheduling: Achieving Equitable Participant and Speaker Satisfaction*. 2020. arXiv: 2010.14624 [cs.SI].

[Pie+20] Benjamin C. Pierce, Michael Hicks, Crista Lopes, and Jens Palsberg. "Conferences in an Era of Expensive Carbon". In: *Commun. ACM*. Vol. 63. 3. New York, NY, USA: Association for Computing Machinery, Feb. 2020, pp. 35–37. DOI: 10.1145/3380445. URL: https://doi.org/10.1145/3380445.

[PMR19] Ben Purvis, Yong Mao, and Darren Robinson. "Three pillars of sustainability: in search of conceptual origins". In: vol. 14. 3. Springer, 2019, pp. 681–695. URL: https://doi.org/10.1007/s11625-018-0627-5.

[Pop18] Celia Popovic. *Learning from Academic Conferences*. Leiden, Niederlande: Brill, 2018. ISBN: 978-90-04-37301-3. DOI: https://doi.org/10.1163/9789004373013. URL: https://brill.com/view/title/38687.

[Por+22] Lukasz Porwol, Sara S. Metcalf, J. Bradley Morrison, Soon Ae Chun, and Luis Felipe Luna-Reyes. "Facilitating Virtual Conferences: Reflections and Lessons Learned in Two Global Communities". In: *Digit. Gov.: Res. Pract.* Vol. 3. 1. International Conference on Digital Government Research, International System Dynamics Conference. New York, NY, USA: Association for Computing Machinery, Mar. 2022. DOI: 10.1145/3494676. URL: https://doi.org/10.1145/3494676.

[Pra+20] Dinesh K. Pradhan, Joyita Chakraborty, Prasenjit Choudhary, and Subrata Nandi. "An automated conflict of interest based greedy approach for conference paper assignment system". In: *Journal of Informetrics*. Vol. 14. 2. Elsevier, 2020, p. 101022. DOI: https://doi.org/10.1016/j.joi.2020.101022. URL: https://www.sciencedirect.com/science/article/pii/S1751157719301373.

[Pre00] L. Prechelt. "An empirical comparison of seven programming languages". In: vol. 33. 10. 2000, pp. 23–29. DOI: 10.1109/2.876288.

[RMA18] Bill Rogers, Masood Masoodian, and Mark Apperley. "A Virtual Cocktail Party: Supporting Informal Social Interactions in a Virtual Conference". In: *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*. AVI '18. Castiglione della Pescaia, Grosseto, Italy: Association for Computing Machinery, 2018. ISBN: 9781450356169. DOI: 10.1145/3206505.3206569. URL: https://doi.org/10.1145/3206505.3206569.

[Roc22] Hugo Filipe Oliveira Rocha. *Practical Event-Driven Microservices Architecture. Building Sustainable and Highly Scalable Event-Driven Microservices*. Apress Berkeley, CA, 2022. ISBN: 978-1-4842-7467-5. URL: https://doi.org/10.1007/978-1-4842-7468-2.

[Roo+20] Goedele Roos, Julianna Oláh, Rebecca Ingle, Rika Kobayashi, and Milica Feldt. "Online conferences – Towards a new (virtual) reality". In: *Computational and Theoretical Chemistry*. Vol. 1189. Virtual Winter School on Computational Chemistry. Elsevier, 2020, p. 112975. DOI: https://doi.org/10.1016/j.comptc.2020.112975. URL: https://www.sciencedirect.com/science/article/pii/S2210271X20302759.

[RS20] Noa Roy-Hubara and Arnon Sturm. "Design methods for the new database era: a systematic literature review". In: vol. 19. 2. Springer, 2020, pp. 297–312. URL: https://doi.org/10.1007/s10270-019-00739-8.

[RSP21] Daniel Reinhard, Mark C. Stafford, and Troy C. Payne. "COVID-19 and Academia: Considering the Future of Academic Conferencing". In: *Journal of Criminal Justice Education*. Vol. 0. 0. Routledge, 2021, pp. 1–15. DOI: 10.1080/10511253.2020.1871047. eprint: https://doi.org/10.1080/10511253.2020.1871047. URL: https://doi.org/10.1080/10511253.2020.1871047.

[Sal+20] Mario Salinas, Daniela Giorgi, Federico Ponchio, and Paolo Cignoni. "ReviewerNet: A visualization platform for the selection of academic reviewers". In: *Computers & Graphics*. Vol. 89. Elsevier, 2020, pp. 77–87. DOI: https://doi.org/10.1016/j.cag.2020.04.006. URL: https://www.sciencedirect.com/science/article/pii/S0097849320300443.

[San+20] Karin Sanders et al. "Why academics attend conferences? An extended career self-management framework". In: *Human Resource Management Review*. Elsevier, 2020, p. 100793. DOI: https://doi.org/10.1016/j.hrmr.2020.100793. URL: https://www.sciencedirect.com/science/article/pii/S1053482220300668.

[Sar+21] Sarvenaz Sarabipour et al. "Evaluating features of scientific conferences: A call for improvements". In: *bioRxiv*. Over 260 national and international academic meetings in various disciplines. bioRxiv, 2021. DOI: 10.1101/2020.04.02.022079. eprint: https://www.biorxiv.org/content/early/2021/01/18/2020.04.02.022079.full.pdf. URL: https://www.biorxiv.org/content/early/2021/01/18/2020.04.02.022079.

[Sea22] Kent Seamons. "Privacy-Enhancing Technologies". In: *Modern Socio-Technical Perspectives on Privacy*. Ed. by Bart P. Knijnenburg et al. Cham: Springer International Publishing, 2022, pp. 149–170. ISBN: 978-3-030-82786-1. DOI: 10.1007/978-3-030-82786-1_8. URL: https://doi.org/10.1007/978-3-030-82786-1_8.

[SFS19] Maria José Sá, Carlos Miguel Ferreira, and Sandro Serpa. "Virtual and Face-To-Face Academic Conferences: Comparison and Potentials". In: vol. 9. 2. May 2019, p. 35. URL: http://www.richtmann.org/journal/index.php/jesr/article/view/10431.

[SRM21] Jaeyoon Song, Christoph Riedl, and Thomas W. Malone. "Online Mingling: Supporting Ad Hoc, Private Conversations at Virtual Conferences". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445776. URL: https://doi.org/10.1145/3411764.3445776.

[TB11] Andrew Twine and Irwin Brown. "Evaluating Web Conferencing Tool Effectiveness". In: *Proceedings of the South African Institute of Computer Scientists and Information Technologists Conference on Knowledge, Innovation and Leadership in a Diverse, Multidisciplinary Environment*. SAICSIT '11. Cape Town, South Africa: Association for Computing Machinery, 2011, pp. 239–248. ISBN: 9781450308786. DOI: 10.1145/2072221.2072249. URL: https://doi.org/10.1145/2072221.2072249.

[TS22] Otto Toivanen and Yossi Spiegel. "From conference submission to publication and citations: Evidence from the EARIE conference". In: (2022). URL: https://repec.cepr.org/repec/cpr/ceprdp/DP16990.pdf.

[TWL20] Md Tahsin Tausif, RJ Weaver, and Sang Won Lee. "Towards Enabling Eye Contact and Perspective Control in Video Conference". In: *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST '20 Adjunct. Virtual Event, USA: Association for Computing Machinery, 2020, pp. 96–98. ISBN: 9781450375153. DOI: 10.1145/3379350.3416197. URL: https://doi.org/10.1145/3379350.3416197.

[UD21] Julia Utzerath and Rhea Dennis. "Numbers and statistics: data and cyber breaches under the General Data Protection Regulation". In: vol. 2. 2. Dec. 2021, pp. 339–348. DOI: 10.1365/s43439-021-00041-8. URL: https://doi.org/10.1365/s43439-021-00041-8.

[Uni] Deque University. *Deque Web Accessibility Checklist*. Online. https://media.dequeuniversity.com/public/en/docs/deque_web_accessibility_checklist.pdf (accessed: September 2022).

[Vah+21] Sahar Vahdati et al. "A comprehensive quality assessment framework for scientific events". In: vol. 126. 1. Jan. 2021, pp. 641–682. DOI: 10.1007/s11192-020-03758-1. URL: https://doi.org/10.1007/s11192-020-03758-1.

[Ven+18] Colin C. Venters et al. "Software sustainability: Research and practice from a software architecture viewpoint". In: vol. 138. 2018, pp. 174–188. DOI: https://doi.org/10.1016/j.jss.2017.12.026. URL: https://www.sciencedirect.com/science/article/pii/S0164121217303072.

[Ver22] Verizon. *Leveraging Video Conferencing Analytics and Insights to Measure Performance*. 2022. URL: https://www.bluejeans.com/resources/guides/measuring-video-conference-performance.

[Vig20] Giuliana Viglione. *A year without conferences? How the coronavirus pandemic could change research*. 2020. URL: https://www.nature.com/articles/d41586-020-00786-y.

[VK22] Virgilio Viegas and Oben Kuyucu. "Security Metrics". In: *IT Security Controls: A Guide to Corporate Standards and Frameworks*. Berkeley, CA: Apress, 2022, pp. 221–244. ISBN: 978-1-4842-7799-7. DOI: 10.1007/978-1-4842-7799-7_7. URL: https://doi.org/10.1007/978-1-4842-7799-7_7.

[WB13] Karl E Wiegers and Joy Beatty. *Software Requirements 3*. USA: Microsoft Press, 2013. ISBN: 978-0-7356-7966-5. DOI: 10.5555/2543993.

[Wil+21] Kristin Williams, Bridget Kane, Chris Clarke, and Kelly Widdicks. "The Environmental Impact of Research Communities: Insights from Conference Sustainability Chairs". In: *XRDS*. Vol. 27. 4. New York, NY, USA: Association for Computing Machinery, June 2021, pp. 46–51. DOI: 10.1145/3468216. URL: https://doi.org/10.1145/3468216.

[Woo+21] Phillip Woodruff, Christopher J.D. Wallis, Peter Albers, and Zachary Klaassen. "Virtual Conferences and the COVID-19 Pandemic: Are We Missing Out with an Online Only Platform?" In: *European Urology*. Elsevier, 2021. DOI: https://doi.org/10.1016/j.eururo.2021.03.019. URL: https://www.sciencedirect.com/science/article/pii/S0302283821002256.

[WP22] Yang Wang and Charlotte Emily Price. "Accessible Privacy". In: *Modern Socio-Technical Perspectives on Privacy*. Ed. by Bart P. Knijnenburg et al. Cham: Springer International Publishing, 2022, pp. 293–313. ISBN: 978-3-030-82786-1. DOI: 10.1007/978-3-030-82786-1_13. URL: https://doi.org/10.1007/978-3-030-82786-1_13.

[Zha+21] Yazhou Zhang et al. "Learning interaction dynamics with an interactive LSTM for conversational sentiment analysis". In: *Neural Networks* 133 (2021), pp. 40–56. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2020.10.001. URL: https://www.sciencedirect.com/science/article/pii/S0893608020303567.

JOHANNES KEPLER
UNIVERSITY LINZ

# Appendix

This appendix contains additional material that supplements the implementation work of this thesis.

## 1 Sourcecode

The sourcecode can be found on Github via `https://github.com/lmhuber/masterproject`.

## 2 Full Class Diagram

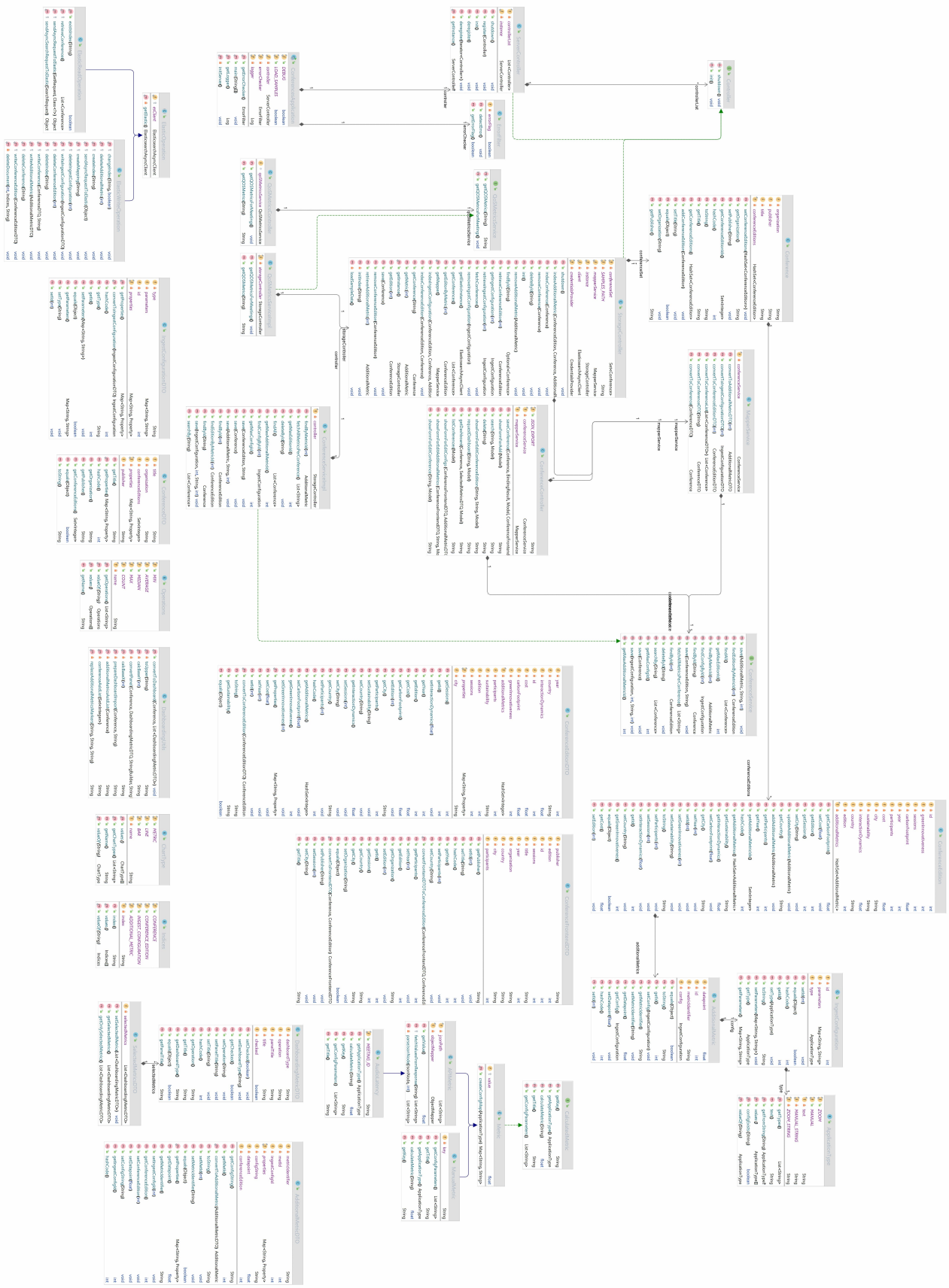


**Figure 1:** Class diagram of the reference framework