Internal and external social dimensions of linguistic legacy materials

The case of Kraasna South Estonian

Dissertation von Tobias Weber

München 2023

Internal and external social dimensions of linguistic legacy materials

The case of Kraasna South Estonian

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
der Ludwig-Maximilians-Universität München

vorgelegt von
Tobias Weber

aus
Mainz

2023

Erstgutachterin:	Prof. Dr. Elena Skribnik
Zweitgutachterin:	Prof. Dr. Ksenia Shagal
Tag der mündlichen Prüfung:	15.05.2023

Für meine Eltern – Marion und Wilhelm

Contents

Zusammenfassung **vii**

1 Introduction **1**

2 Background and Motivation **6**
2.1 Kraasna 8
2.1.1 Artefacts of past documentation projects 9
2.1.2 The Kraasna community 11
2.1.3 The Kraasna variety 11
2.2 Linguistic legacy materials 13

3 A framework for linguistic legacy materials **15**
3.1 Artefacts 17
3.2 Knowledge 23
3.3 Archives 27
3.4 Using the framework 30

4 Outlook **36**

5 Published works **39**
5.1 Paper I 39
5.2 Paper II 39
5.3 Paper III 39
5.4 Paper IV 40

Bibliography **41**

Acknowledgements **53**

List of Figures

2.1 Cyclical model of archival work 8

3.1 A network of data artefacts and scientific texts. 34

List of Tables

1.1 Typology of representable aspects of science. 4

3.1 Domains of knowledge creation 16
3.2 Ownership and attribution 32
3.3 Anthroponyms in the framework 33

Zusammenfassung

Linguistische Nachlassmaterialien sind Artefakte der modernen Wissensgesellschaft, die Einblicke in die Sprach- und die Wissenschaftspraxis früherer Generationen geben. Für die Rekonstruktion ihrer Dokumentationskontexte fungieren sie oft als primäre Quellen, für Beteiligte durch ihre mnemonische Funktion oder in Ermangelung dieser als Zeitzeugnisse. Die vorliegende Arbeit beschäftigt sich mit der Dynamik des Wissensverlusts am Beispiel linguistischer Nachlassmaterialien, sowie der benötigten Rekonstruktion und Aufarbeitung im Umgang mit diesen Artefakten. Die Arbeit stellt hierzu einen hermeneutischen Beschreibungsrahmen vor, in dem der menschliche Faktor als soziale Dimension zwischen Wissen und Artefakt vermittelt. Dieses Modell unterscheidet zwischen einer internen und externen Referenzebene und den mental-abstrakten und materiellen Repräsentationen von Wissen. Die Kombination dieser Ebenen spannt vier Beschreibungsfelder auf, wobei der Transfer von Wissen zwischen diesen Bereichen durch menschliche Handlungen zu Erkenntnisgewinnen führt.

Im Fallbeispiel der Nachlassmaterialien des südestnischen Kraasna-Dialekts liegen ausschließlich Artefakte von drei größeren und zwei kleineren philologischen Dokumentationsprojekten vor, die zwischen 1849 und 1968 entstanden sind. In diese Zeit fällt der Übergang von einer aktiven alltäglichen Sprachnutzung zum Vergessen des linguistischen Erbes der um die russische Stadt Krasnogorodsk (Oblast' Pskov) ansässigen Sprachgemeinschaft. Die Zielsetzung bei der Bearbeitung dieser Nachlassmaterialien war die Aufarbeitung dieser Materialien, die Beleuchtung der diversen gesellschaftlichen wie wissenschaftlichen Kontexte zu dieser Zeit, sowie die Untersuchung der Rezeptiongeschichte zum Kraasna-Dialekt und seinen Sprecher:innen. Im Zuge dieser Arbeit sind neue Einblicke in die Wissenschaftsgeschichte entstanden, insbesondere bei den wiederentdeckten Tonaufnahmen aus dem Jahr 1914, welche im Rahmen dieser Dissertation neu transkribiert und erstmalig analysiert wurden.

Artikel II widmet sich der linguistischen Beschreibung der acht Phonographenwalzen, welche eine unschätzbare Quelle für die linguistische Beschreibung des Kraasna-Dialekts darstellen. Die Analyse bezieht dabei auch bereits veröffentlichte Beschreibungen ein und gelangt zu einer umfassenden Darstellung der Sprache, wobei sich verschiedene linguistische Annahmen bestätigen oder widerlegen lassen. **Artikel I** resümiert die notwendige philologische Rekonstruktion, um die diversen Nachlassmaterialien in ihren Kontexten der Erstellung und Rezeption zu begreifen. Hierbei ist eine reflexive Haltung vonnöten, die nicht nur Positionalität und menschlichen Einfluss in den Materialien untersucht, son-

dern auch die eigenen Spuren des Kurators transparent hält. **Artikel III** überträgt das Konzept der Kuratierung auf die wissenschaftliche Tätigkeit und Ausbildung. Diese Arbeit bedarf interdisziplinären Austauschs, einer holistischen geisteswissenschaftlichen Perspektive und vielseitigen Karrierewegen in der Aufarbeitung sprachlicher Daten und Dokumentationserzeugnisse. In **Artikel IV** stehen Archivinfrastrukturen im Fokus, die auf die Bedürfnisse von linguistischen Nachlassmaterialien und deren Aufarbeitung zugeschnitten sind. Die empfohlene partizipative Archivform kann die Gedächtnisinstitutionen zum Ort des Austauschs und der Verhandlung zwischen verschiedenen Interessengruppen erheben, bei der durch die Nachlassmaterialien auch eine historische Aussöhnung und ein Dialog zwischen Generationen ermöglicht wird.

Die vorliegende Arbeit verknüpft die Thematik der Nachlassmaterialien mit wissenschaftstheoretischen und -geschichtlichen Ansätzen und stellt eine metawissenschaftliche Perspektive auf diese Artefakte der linguistischen und anthropologischen Forschungsarbeit dar. Für die estnische Dialektologie bietet die Arbeit eine umfassende Aufarbeitung der Forschungsgeschichte und der diversen Nachlassmaterialien, welche in den südestnischen Kontext eingeordnet werden. Darüber hinaus bietet die Fallstudie einen Referenzpunkt für die Reflexion und Theoriebildung im Bereich Sprachdokumentation bzw. Metadokumentation, sowie Anhaltspunkte für die linguistische Arbeit mit Nachlassmaterialien. Des Weiteren illustriert die Arbeit Ansätze und Entwicklungsmöglichkeiten für moderne partizipative Infrastrukturen in Wissens- und Gedächtnisorganisationen, die auf die Besonderheiten im Umgang mit Nachlassmaterialien ausgerichtet sind. Die Aufarbeitung und Kuratierung dieser Artefakte verdient größere Aufmerksamkeit und die gleiche Anerkennung wie andere Forschungsarbeiten, da sie einen wertvollen Beitrag zur vielfältigen Beschreibung und selbstreflektierten Wissenschaft bildet.

Der Umgang mit Nachlassmaterialien stellt einen Diskurs dar, bei dem Wissen verhandelt und konstruiert wird. Im Hinblick auf diese soziale Konstruktion von Erkenntnissen von der Dokumentation über die Aufbereitung bis hin zu modernen Veröffentlichungen ist ein Fokus auf die sozialen Dimensionen wissenschaftlicher Artefakte und damit verbundenen Praktiken notwendig. Hierbei geht es nicht nur um einen kritischen und selbstreflektierten Blick, sondern auch um Respekt und Anerkennung für die menschliche Arbeit in der Wissensgenese. Heutige Generationen können aus der Arbeit vergangener Generationen lernen und auf dieselbe Weise nachfolgenden Wissenschaffenden aus Gesellschaft, Bildungswesen und Verwaltung helfen, ihre eigenen Perspektiven und Prozesse in der Dokumentation und Kommunikation von Wissen nahezubringen. Transparenz bildet ein verbindendes Element zwischen den Generationen und ermöglicht es, Kontexte zu rekonstruieren und Prozesse nachzuvollziehen.

Chapter 1

Introduction

Knowledge is inseparably linked to humans, their cognitive processes, individual and collective memory, and traditions of generating and disseminating information among each other. It is not identical to objective truth (*veritas* in a Heideggerian sense opposed to *aletheia*), which most modern academic traditions would claim exists outside of human perception and description[1]. It is rather a contextualised, mental representation of reality. Yet, as researchers, we aim to generate and spread knowledge that resembles and overlaps with an objective and empirically grounded truth, as a part of our professional self-conception[2]. If the natural equilibrium contains exclusively abstract 'truth' about the real world, without human interference, then the generation of knowledge introduces entropy to the system through individual accounts approximating this true state of existence[3], yet without covering or describing the entirety of it. Thus, all human processes generating knowledge in an attempt at describing existence, including religion and science, face challenges and changes to the beliefs and accepted truths. These arise as more and different knowledge is produced and further accounts of reality are included in the respective ontologies. Therefore, the

[1]Some philosophers in phenomenology and metaphysics equate perception with ontic truth (Husserl, 2002 [1913]) while others establish a link that does not preclude existence outside of human thought and insight(Heidegger, 2018 [1927]; Lohmar, 1997). This thesis follows scholars who seek to establish a unique view of existence from the viewpoint of the humanities, without conflating the description with the originary (Dilthey, 1990 [1883]; Derrida, 1973 [1967]). As Gadamer (1975 [1960]) outlines the humanities' aim in understanding rather than following the natural sciences in their search for a firm truth as explicit facts, knowledge and meaning in these disciplines always bear a subjective element of cognitive processes rather than constituting a tangible object.

[2]Knowledge in this sense implies that there are not just mental states of the memory, but posits that these perceptions and beliefs have a justification for the derived knowledge, which can be shared with others. Data, as means of disseminating an observable basis for knowledge creation, are frequently seen as a way to establish the foundation for scientific discovery. Knowledge is subsequently also linked to the interpretation of data, i.e. the contextualisation of information.

[3]On an absolute, universal scale, existence is not affected by human actions of knowledge generation. However, on a global level and in areas of human life, our actions including scientific enquiry shape our planet noticeably, e.g. climate change/action, constructions and infrastructure, technology. Especially in the humanities and social sciences, there is a direct link between analyses or proliferated thought and the design of our cultural, intellectual, and societal lives.

only state of homoeostasis is the state without human knowledge, as constant changes, additions, and omissions in our ontological descriptions preclude absolute accounts of reality. In thermodynamics, the discipline from which information theory has adopted the terms 'entropy' and 'homoeostasis', this is closely linked to concepts of causality and time. Likewise, these concepts affect the state of human knowledge: On the one hand, knowledge does not emerge naturally but in a causal relationship to perception and theory building, on the other hand, the relation between different events in this process (i.e. time) leaves its traces.

This thesis investigates the link between particular processes of knowledge generation and dissemination in a small area of the humanities, namely linguistics and its neighbouring disciplines in anthropology and the social sciences. In this discipline, in the tens, maybe a hundred, of thousand years that humans have used a system similar to our modern notion of language (Nichols, 1998; Perreault & Mathew, 2012), only observations since the Bronze Age have been recorded and preserved, while the scientific traditions of analysis and description are even younger. These analyses are based on the abstract understanding of symbols not as a direct representation of an object, as in parietal art, and not even as abstractions like logographs or words, but as the depiction of the abstract system itself, e.g. phonetic transcriptions. These are inherent in primary and secondary data (Lehmann, 2004; Himmelmann, 2012) that linguists use to generate knowledge by description, theorisation, and inference over these accounts of language use. In many of these instances, the object of description and focal point of theory building is documented language use that is bound to a medium of a textual, visual, or auditory nature. The actual observation of the underlying speech event and the speakers' knowledge at the point of recording may have been lost through the course of time; there are many undeciphered scripts and artefacts bearing symbols, the meaning and interpretation of which is unknown to modern scientists[4]. Despite the primacy of speech over its written representations in linguistics, 'graphism' defines the boundaries of investigation beyond the invention of audio recordings (Leroi-Gourhan, 1993; Fynsk, 2003) and with it a critical perspective of 'scriptism' (Harris, 1980) focused on the medial representations. In this respect, the *artefact* as a medium containing language data – be it a Sumerian clay tablet, an early modern ethnographer's field diary, or a present-day audio recording – is always subjected to the effects of time, i.e. events between its creation and modern-day use. For language data and linguistic legacy materials, it is important to consider that these contain different contextual layers based on the mental representations of their creators, different standards like transcription rules or orthographies, and medial restriction. Who wrote what, where and when, and with which means, to what end are contextual clues that shape the artefacts.

In this thesis, *legacification* shall be understood as a gradual shift from primacy of memory, and with it the spoken word, to the primacy of a technologically mediated, artefactual representation of the event. Thus, it describes a process of human memory in its

[4]Whether or not these inscriptions and symbols are considered as a script is an important debate but would distract from the relevant points about written accounts in linguistic research (e.g. Derrida, 1997). We will return to processes of writing in chapter 3.

relationship to abstractions and representations of knowledge. In the definition of linguistic legacy materials, the artefact is more important in understanding the underlying speech event than human memory and knowledge. This knowledge might be inaccessible as a result of time between recording and today[5]. Consequently, this definition is not identical to concepts presented as "legacification" in computer science or information technology, where it is used synonymously to 'technological obsolescence'. While so-called legacy formats confront the user with similar accessibility issues, the relatively recent advent of most complex information technology means that we often have ways to restore access to data without the need to infer historical knowledge from the artefact (Aristar-Dry, 2009). Human knowledge and any observations preserved in our memory will be distorted, if not lost, through the course of time. Yet, artefacts as bearers of data, information, or even knowledge also face effects of time, which will be a focus of the analyses in chapter 5. In order to access the knowledge contained in them and to generate new knowledge based on them it is necessary to interpret, contextualise, and understand changes that occurred to the artefacts in their respective (physical) contexts, which includes both physical deterioration and historic processes of meaning-making. To capture this bipartite nature of legacification, the framework in chapter 3 will focus on processes related to artefacts as well as to humans in their interaction with artefacts.

Five cognitive processes are relevant in the discussion of legacy materials and the human role in them: meaning-making, knowing, forgetting, remembering, reconciling. First, knowledge is generated through meaning-making processes that aim to understand and offer an explanation of the world around the human observer. Second, these ideas are processed as active knowledge by the individual but can also be shared through oral tradition or artefacts, in a stage of knowing and knowledge dissemination (as a form of negotiating meaning). As discussed above, with increasing (temporal, spatial, cultural) distance from the original event and context of knowledge generation, the perspective on what is known or considered knowledge changes. The third stage thus comprises two processes that filter the initially generated knowledge, namely forgetting and remembering, which may undulate over time. Although in our societal judgement forgetting is negatively connotated, this should be seen as a measure of relevance and, while not remembering important pieces of information should be perceived as an issue, frees capacities for new knowledge that can be contextualised vis-à-vis remembered information. For example, most people will have forgotten all details about a geocentric model and its religious justifications, but will remember that such approaches once existed before being superseded with modern cosmological models. In this final stage, formerly held knowledge is reconciled with new insights, thereby creating a loop that connects the 'final' stage with new cycles of knowledge generation. In all stages, we need to investigate the human mind in relation to its artefactual

[5]There is no firm threshold for an artefact to be considered legacy material; the process is gradual, starting at its creation. This is also reflected in the different time spans of data and privacy protection in legal texts around the world, which may outlast the human life and, thereby, their biological capacity of memory (Austin, 2010; Assmann, 2011; Banta, 2016; Buitelaar, 2017). This memory can be lost in a gradual process of forgetting or, occasionally, in abrupt ways if a researcher cannot conclude a research project. Therefore, recording metadata and information about an interview is advisable as soon as possible.

representations; there are two dimensions to both the material or artefactual level and the abstract knowledge level. These will be discussed as *internal* and *external* dimensions, i.e. artefacts or knowledge *as* a system versus *in* a system. In other terms, what can be found in the same archive folder or box, the same archive, project collection, book, or report; many different concepts can be applied to cover the artefactual and real-world systems in which our descriptions of the internal dimensions apply. Yet, as soon as we need to move beyond the text or artefact, beyond an (academic) approach or tradition, and consider how each of our microscopic entities relate to externally posited knowledge or other artefacts, the analysis affects the external dimension in a macroscopic view.

In contrast to mere objects, linguistic legacy materials need to be approached with a view to the contexts in which they were created, including their relationship with knowledge generation and the human role in that process. As they are both artefacts as well as bearers of data, their dual nature needs to be considered. Linguistic legacy materials are more than containers for data, they are the basis for reconciling generations of academic practices and language users. Language data exist in the grey area between ontology and epistemology because their intended purpose within knowledge generation links them to epistemic processes. They are expected to represent language use on an ontological level, i.e. a linguist will expect language data to accurately depict language use as (was) observable in reality. At the same time, the narratives, texts, or words contained within bear meaning in themselves, describing the speakers' relationship towards their surroundings in a meaning-making sense. Through these texts, it is also possible to learn more about what was known at the time or recording – social scientists may consider this level more important than structures and functions of the language (e.g. Vogel, 1989; Brázdil et al., 2005). As users, we are thus moving between the textual and the contextual plane, where the text relates both to textual (i.e. cotext) and real-world contexts. In a conference paper that inspired this theoretical framework, I described the movement from data to publication under the same premise (Table 1.1). In this table, the internal and external dimensions are mapped into two directions, between meso- and macro-level and between the 'concrete' (artefact-internal) and 'abstract' planes. The framework in chapter 3 spans this field differently with a focus on artefacts and humans, *as* or *in* a system.

	Concrete	**Abstract**
macro-level	context	meta-context
meso-level	text or cotext	meta-text or commentary
micro-level	constituent or 'component'	metadata

Table 1.1: Typology of representable aspects of science.
(Weber, 2020b, 228, reproduced with permission from Springer Nature)

Eventually, these typologies and frameworks bear witness to the fact that there are multiple levels of abstraction and representation involved in knowledge generation and the creation of scientific artefacts: Language represents cognitive processes, script represents language, artefacts contain mediated language use (including script or sound waves), digitised versions emulate the physical object, while digital dissemination creates numerous

versions of the virtual object itself (Weber, 2020a). All of these planes have their own relationship to each other and in regard to the ontic reality; none *is* absolute, originary reality but merely represents different parts of it filtered through restrictions of the medium and human choices in its creation. Consequently, this thesis adopts a reflexive stance in applying scientific methods to the 'products' of scientific enquiry. At the core of this meta-scientific endeavour lies the goal to investigate the hermeneutics and philological ties in linguistic legacy materials, in order to create a practical framework for dealing with them. This does not imply that an absolute or definite guideline could be written – the framework contains procedures that were relevant in the case study of the Kraasna linguistic legacy materials and their reconstruction, which may be adapted to other contexts. The following chapter introduces the peculiarities of this case and presents necessary background information.

Chapter 2

Background and Motivation

In times of seemingly unlimited, ubiquitous data and amounts of digital information increasing daily, one may ask why the study of a language form that only exists in archival records and deprecated media formats is of interest to modern science. Undoubtedly, the concerns and issues of present-day language communities, especially those under pressure from languages of wider communication, should be a priority for any socially conscious linguist. At the same time, the formation of this awareness arises from a critical view of present and past practices in the interaction with language communities. It depends on the reflection of the linguist's stance within the processes of documenting, communicating, and creating knowledge about and for these communities, including information on language practices and sociolinguistic situations. The case study of the South Estonian Kraasna variety is, to some extent, an illustrative example of how past and present can be reconciled in abstract knowledge and associated material artefacts. Similarly, the papers in this dissertation display a research trajectory that aligns with the time of my studies of this variety. It begins with me first learning about it from a text collection (Mets et al.) I obtained right after its publication in 2014 while studying abroad in Estonia, through archival work for a Bachelor and Master thesis, to the work at hand. Initially, I approached the Kraasna variety from the perspective of descriptive linguistics, quickly realising that a broader approach would be warranted, if not imperative. Through expansion into digital humanities, documentary linguistics, and ethnography, the framework of philology was my preferred choice when I submitted Paper I at the start of my doctoral studies. Since then, I have broadened the scope further and adopted a perspective from hermeneutics and philosophy of science that informs this dissertation at the time of writing in late 2022. Thus, it captures a state of knowledge at this point in time which, as will be discussed in chapter 3, is not impervious to change even if this text does not change beyond publication.

The analysis presented in the chapters of this dissertation and the papers contained within are not merely anecdotal, nor do they form an autoethnography (Poulos, 2021). Through critical reflection it has become obvious that researchers are linked to their research and need to communicate their decisions if future academics are to understand and use their work appropriately. This follows similar developments in exploring the researcher role under the idea of 'me-search' (Gardner et al., 2017) or 'we-search' that emphasises

collective self-conceptualisations (Douglas, 2017). Certainly, researcher positionality can influence public trust by reinforcing positive or negative prejudices against the research if the investigators declare a position (Altenmüller et al., 2021). The latter part is more critical for polarising research topics, whereas academic communities are increasingly adopting the concept of 'we-search' to investigate biases and narratives within knowledge and memory institutions (e.g. libraries, see Winberry & Gray, 2022). This is also a prerequisite for the philological work I conducted, where transparency is important for future examination and discussion of my work, while I approached the task at hand with a similarly mindful perspective of subjective narratives embedded into the artefacts. As Gurd (2015) reminds us: 'One cannot describe past philologies as erroneous without acknowledging the likelihood that one's own certainties will one day fall under a similarly critical eye'. This does not preclude their analysis and critique, nor does it issue a *carte blanche* to researchers that 'errors' or scientific misconduct will be tolerated based on the subjective perspective in their work. In my view, researchers need to adopt a reflexive stance, and accept the responsibility for their actions and scholarly communications without hiding behind an anonymous community and its 'objective' research agenda[1]. Reflexivity is always a part of linguistics (e.g. Wayt, 2021, 141), and even if there is some level of insecurity to disclose positions or motivations underlying our research, this is part of our professional self-conceptualisation (Jackson, 1990).

The artefacts we produce as members of the academic community, data sets or publications, need to be examined for embedded narratives that allow for an analysis of the contexts in which they were generated. The circulation of knowledge and artefacts is thus always tied to human actions within relevant contexts, e.g. institutions of education, archives, communities within society. By placing emphasis on the human role, we can understand the 'social lives of legacy materials' (Dobrin & Schwartz, 2021). Figure 2.1 illustrates processes of archival work in respect to communities, with a central position afforded to past and present language communities and researchers. The illustration shows that they are as much part of the contemporary research and enter the discourse through the artefacts they left behind as the present day researchers and communities. At each stage, we need to consult these communities and reconcile our knowledge with theirs; the analysis can spark new directions or ideas for research and development, e.g. of educational programmes (Paper III) or archival infrastructures (Paper IV). My work on Kraasna started in the selection of a research question that was influenced by the publication of the dialect text collection (i.e. the connection to a previous cycle of evaluation, curation, and dissemination by colleagues) and conversations I had with supervisors and fellow students. This was followed by lengthy evaluation and curation tasks, as outlined in Paper I and Paper II. This work is now sparking new conversations and areas for future research

[1]This dissertation is written in first person to differentiate between others' and my own position or to indicate where I have taken the agency to make an informed decision in my research (Weber & Klee, 2020). The passive stance of remaining anonymous in the crowd, which can be observed in some publications, is an active decision trying to separate human perception from the generation of knowledge. The active questioning and understanding requires a different stance than the impersonal 'Man' (Heidegger, 2018 [1927]).

activity, thereby continuing the cycle.

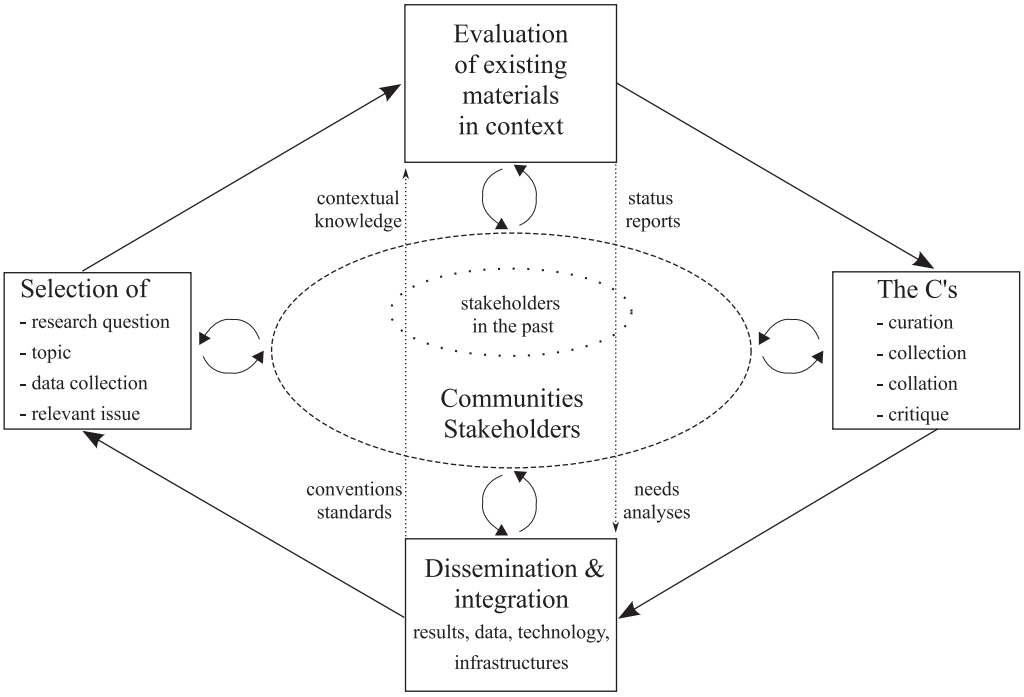

Figure 2.1: Cyclical model of archival work with communities at its centre. (Paper IV, 2022 ©Emerald Publishing Limited all rights reserved)

2.1 Kraasna

The South Estonian Kraasna community existed as a linguistic enclave among a Russian majority around the city now called Krasnogorodsk in the Pskov Oblast'. This community existed until the mid-20th century; the last proficient speakers passed away in the 1930s, leaving rememberers of the language and South Estonian heritage that were still interviewed in 1968[2]. Therefore, our knowledge stems exclusively from historical sources, including publications like Oskar Kallas' ethnographic monograph *Kraasna maarahvas* (1903) and archival records. Although contextual information about the Kraasna Estonians and their language is contained in these publications as well as in each paper of this dissertation, the following will provide a brief summary of artefacts, the community, and its language.

[2]The recordings are held in the University of Tartu Archives of Estonian Dialects and Kindred Language – recordings F1223-02 and F1223-03 – where the consultants mention single words in the Kraasna variety.

2.1.1 Artefacts of past documentation projects

The main bulk of archival data on the Kraasna community and the South Estonian variety they used stems from two researchers, Oskar Kallas and Heikki Ojansuu[3]. These materials include a monograph (Kallas, 1903), as well as a number of articles (e.g. Kallas, 1902; Ojansuu, 1912), field notes with ethnographic and linguistic information, and phonograph recordings from Ojansuu's 1914 fieldwork (Paper II). There are multiple versions of several archival records, where Ojansuu's data had been deposited in Estonia and copied for an archive in Finland and vice versa. This leads to some inconsistencies in the linguistic data which might depict the language differently depending on the underlying data sets. In order to be able to work with these data, I needed to reconstruct relationships between versions and apply methods of philological reconstruction to the data. For understanding the contexts before and after the core period of fieldwork, i.e. Kallas and Ojansuu, there are some data included in an exchange of letters between scholars Adolph Johann Brandt and Friedrich Reinhold Kreutzwald from 1849 (Ernits, 2012), as well as the field diaries of Paulopriit Voolaine who conducted several ethnographic and folkloristic fieldwork trips in the area between 1952 and the 1970s. During latter trips, colleagues and students were also involved in fieldwork, which likely led to the 1968 recordings mentioned above. Of a secondary nature are numerous descriptions and publications on the Kraasna community or its language which are either published or archived. Among those, I would like to emphasise that, despite its peripheral position in Estonian dialectology, a cohort of students at the University of Tartu between 1937 and 1940 (Palgi; Kukk; Nigol) had been actively working with the Kraasna data, forming an intermediary stage between the documentation by Kallas and Ojansuu (1901-1914), and later fieldwork by Voolaine (from 1952). These student papers are especially important, as they fall into a time when the anonymous author of manuscript ES MT 224 interviewed Mrs Ojansuu in Helsinki in January 1938 – the notes are included in the appendix to ES MT 224 which inspired me to search for and rediscover the 'lost' phonograph recordings. It is possible that this interview was part of the preparation for one of these student papers, either by the authors mentioned or an unknown student.

Apart from the scattered artefacts and issues concerning versioning, each artefact is also imbued with the fieldworker's positionality. Kallas notably recomposed the data from his notepads, either by combining phrases or words from different pages or by altering the morphological form to fit the Estonian prose of his monograph. While this might not render 'wrong' utterances in the Kraasna variety or constitute phrases Kallas had not heard during his fieldwork, he did not copy the data as inscribed into his notepads directly into his monograph. A similar issue of filtering and linguistic editing can be observed in Ojansuu's data: One source of divergent manuscripts are the repeated copying and editing that has led to a variety of versions (AES 202, ES MT 224, *Estonica* I-V, Mets et al. 2014, Paper II), the other issue lies in the ephemeral nature of spoken language which, despite recorded onto the phonograph cylinder, is limited by the boundaries of the physical medium. I believe that it was not possible for Ojansuu to make perfect transcriptions of

[3]A detailed description of the archival records is included in Paper I.

his recordings, as the repeated playing would have damaged the cylinders and decreased their use in the archive (see Paper II). Despite this mitigating factor, Ojansuu also made decisions in the transcription of Kraasna data, as to what warranted transcription. In several cases, repetition or self-correction have been removed from the manuscripts that appear to be significantly altered in terms of syntax. In a critical view, the removal of a Kraasna 'style of narration' as a feature of the spoken narratives by Ojansuu presents a case of 'discursive discrimination' (Kroskrity, 2015) against traditional verbal art. However, for both Kallas and Ojansuu, there is 'no neutral basis for judging the decisions [they] made, even when they lead the texts to depart significantly from what was recorded' (Dobrin, 2021, 44), as will be discussed in chapter 3. In practice, we as present-day researchers must be very careful when working with any of these data sets and discern the various linguistic and cultural filters that have been added to the data in their artefactual form. Any use without critical reflection would not do justice to the language, the consultants, or the creators of the records.

Due to the various contextual influences visible in the Kraasna artefacts, the work with the legacy materials requires the researcher to understand how they shaped the documentation and communication practices at the time. Considering mother tongue, Kallas was a native Estonian speaker but from the island of Saaremaa which has a different variety than the South Estonian Kraasna he documented. Ojansuu was a native speaker of Finnish, for whom the South Estonian variety would be a foreign language, just as for Brandt or myself. Consequently, for all researchers apart from Voolaine, South Estonian was not their dominant language, opening the possibility for linguistic interference during fieldwork and analysis. While monolingual elicitation (Everett, 2001) or standardised transcription systems (Setälä, 1901) may remove the direct suggestion of forms or pronunciations, the researchers' native languages, their understanding of the target language, and all ideologies attached to these language forms and their representation are inherently present during fieldwork and documentation. Kallas used transcription rules that were inspired by Uralic folklorists at the time, not yet a standardised transcription system but one peculiar to his rendering of the variety. His field notes are kept in German, Russian, and Estonian, at times skipping between languages and orthographic systems, e.g. Du hast nur S- / *ma jooze mäda, sa joozed verd.* Schluß. 'you have only S- / I exude pus, you exude blood. End.' (EFAM Kallas M4) where he intended to inscribe the meaning in German before changing to the Kraasna South Estonian phrase. We may thus ask why Kallas wrote his field notes - are they just serving a mnemonic function for him to revisit the encounter by reconstruction or are the field notes in themselves a communicative artefact for others? The break in the middle of the sentence shows that Kallas aimed to record Kraasna phrases, although they were embedded in an interview he did not transcribe completely. While writing, he was processing the South Estonian variety of his consultant but also taking notes about the content, not just the form, in his auxiliary languages. Similar phenomena can be observed in Voolaine's notes, where Russian was the language of the interview but his comments appear in Russian and Estonian, occasionally with conflated orthographies, requiring not just knowledge of Estonian or Russian but the simultaneous processing of both in his diaries.

2.1.2 The Kraasna community

The main sources for information about the Kraasna community are Kallas' monograph (1903) and Voolaine's diaries. I would argue that the narratives also hold some value in the reconstruction of community life (Weber, 2021b). However, one must be critical of the prompts that might have led to these narratives. We learn from Kallas' monograph that the community was shifting linguistically to Russian and adopting an increasing number of Russian customs that either supplanted or mixed with traditional customs. At the same time, Kallas' descriptions are not neutral or objective: His goal is to reconcile the Kraasna *maarahvas* with its Estonian kin, while also observing the situation in which the 'Estonian' population might accommodate a 'foreign' Russian element to a larger extent than national romantic Kallas would like to see. In consequence, his description of language endangerment and attrition paints the speakers in a negative light, despite their obviously strong command of their South Estonian variety (see Iva, 2015, Paper II). Not only are the Kraasna Estonians caught between two worlds, a South Estonian heritage and their Russian-dominant everyday lives, but Kallas is also caught between two roles: That of the compassionate 'distant relative' to this Estonian people and the role of the scholar in a Western, colonialist tradition that dismisses certain customs (Lukin, 2017). These positions are important to consider when reading Kallas' published and archived communications about the Kraasna Estonians, making critical reading necessary (Marcus & Cushman, 1982).

2.1.3 The Kraasna variety

The Kraasna variety of South Estonian (Glottocode *kraa1234*) has been the focus of several papers, although Kraasna is usually just one variety among many investigated through larger databases or dialect corpora. However, following the increase in interest after the publication of the 2014 dialect text collection (Mets et al.), there has also been an increased number of publications that focus on Kraasna and other linguistic enclaves ('keelesaared'), i.e. Leivu and Lutsi (Ernits, 2018, 2021; Balodis & Pajusalu, 2021; Norvik et al., 2021). In many cases, however, Kraasna is just briefly mentioned in introductory chapters or general handbooks of Estonian dialectology (Pajusalu, 2007, 2022; Pajusalu et al., 2018). These descriptions highlight that Kraasna is typologically not too distinct from the other South Estonian varieties, especially (Eastern) Seto, despite retaining some archaic features that are visible in Kallas' data and, rarely, also in Ojansuu's manuscripts.

My approach to describing the Kraasna variety was following a framework-free description of the newly transcribed phonograph recordings. As these recordings had not been analysed as such and the manuscripts contained inconsistencies, Paper II focused on the eight cylinders of 20 minutes of recording. These recordings stem from Ojansuu's 1914 fieldwork and appear to be the same communicative event transcribed for the manuscript with language use appearing to be natural in a monologic narrative or structured elicitation. I opted to describe the language of the phonograph recordings in respect to established categories in different dialectological publications, spanning phonology, morphology, and

syntax. A full description would have gone beyond the scope of the paper and might not have yielded the desired thoroughness – the Kraasna data are not linguistically consistent within themselves, some narratives in the manuscripts exhibit a stronger Russian influence than others making it difficult to present an accurate description of *the* Kraasna variety in 1914. In addition, the recordings do not offer full accounts of all morphological categories that a full description would aim to deliver, e.g. no abessive or terminative cases were recorded. As these rare forms are also only sporadically contained in the manuscripts, it is difficult to establish a definite form based on the comparatively small Kraasna corpus.

Since the full description is presented in Paper II, the following is a summary of the most relevant or surprising characteristics of the language in Ojansuu's Kraasna recordings. These recordings confirmed the hypothesis that not only *e* and *i* appear as iotated in word-initial position but also *ü*, e.g. *jüldäs* 'it is said' instead of **üldäs*. This phenomenon can also be observed in the 1968 recordings and Voolaine's interview notes and can potentially be explained through Russian influence. A similar influence might have given rise to the conflation of stress and length, whereby a stressed syllable appears lengthened or slightly diphthongised, e.g. *nāɢɬ* 'nail', *kᵘòrv* 'basket'. The latter has been transcribed in the manuscripts and might be idiosyncratic for some particular consultants. The same explanation may be offered for the surprising observation that an unpalatalised, velar *ɬ* can be heard in palatal contexts, e.g. *śə̄ɬ* 'there'. Other than that, palatalisation largely follows the South Estonian typology, yet it is more frequent under Russian influence. Grade can play a role in the palatalisation of liquids in words of the shape #CVi_V(s), #CVi_, or #CVV_i. In the second morphophonological grade of these words, *i* is elided but triggers palatalisation of the liquid or semivowel, e.g. *haìna* 'hay.PAR' but *hāńa*ʔ 'hay.NOM.PL'. This is one of the phenomena regarding elision frequently ascribed to Kraasna. However, while vowel elision or syncope is noticeable in the recordings, it is less strong than implied by the manuscripts. Even in cases of an extreme reduction, a schwa or audible syllable break might remain of the elided vowel.

In terms of morphology and syntax, the translative is recorded less frequently than would be expected. While it occurs in the the form *-st* indicated in the literature, there is a pair of semantically identical phrases – *Ĺät haìgest* and *sā haìge* 'falls ill' – where only the first is marked with the translative suffix. The underlying motivation could not be identified for these examples, although different explanations are offered in Paper IV. Overall, there are some surprising instances of syncretism in the manuscripts, namely in the second person marking in the verbal paradigm; the data from *Estonica* V contain forms such as *annàde* that were used in singular and plural (cf. Norvik et al., 2021, 52), which could not be checked in the recordings. For a holistic view of the Kraasna variety, it is necessary to compare the language of the recordings to the documented language use in the manuscripts and Kallas' data. This is especially relevant for verbal morphology, as the recordings contain predominantly necessitative, inchoative, and impersonal passive constructions instead of personal finite verb forms. These might result from the genre of the narrative or framed by the question in the interview, e.g. a prompt 'how are funerals usually arranged?' might yield a more impersonal or procedural narrative. In the necessitative constructions, the copula is often omitted, either as a characteristic of the spoken language form or through

Russian influence. Overall, and despite Kallas' account of language attrition, the speakers in the phonograph recordings have a good command of the language and can produce a monologue without major pauses or breaks in the narrative. Occasional self-corrections are negligible but demonstrate that the consultants were very aware of their language use, e.g. *pandas pad́a pä- pandas padi pāla* 'the pillow is put underneath the head', where the speaker notices that the impersonal passive form she used would regularly use the object in nominative (*padi*) rather than genitive (*pad́a*). This is a sign of linguistic awareness rather than language attrition.

2.2 Linguistic legacy materials

The preceding discussion has already illustrated some issues that are not unique to the Kraasna legacy materials. Similar observations have been made by different scholars from various linguistic disciplines, which will be the starting point of the theorisation in the next chapter. Before moving on to the framework, I will conclude with some background information on linguistic legacy materials as a type of artefact or label assigned in the literature. Peter K. Austin, who inspired my interest in the theory of linguistic legacy materials, lists different possible issues that can arise when working with legacy materials (Austin, 2017): form, content, context, and stakeholder issues. The latter, like access and rights management in the archive, have not affected my work with Kraasna apart from the identification of speakers. Meanwhile, form, content and context issues are all observable in isolation and interaction – deciphering and re-transcribing the notes; interpreting different linguistic and cultural filters that have been applied by researchers, transcribers, editors, and not least the speakers themselves; unclear and scattered metadata requiring reconstruction and recontextualisation. Blokland et al. (2021) outline five different perspectives they adopted in working with legacy materials: a view on artefactual history; ethnolinguistic embedding; and linguistic, documentary, and technological perspectives. While each of these areas has a body of literature focused on linguistic, technological, or archival issues, a theory of linguistic legacy materials needs to go beyond finding solutions to microscopic issues like the automated conversion between transcription systems or the annotation of linguistic examples. The framework presented in this dissertation aims to discuss linguistic legacy materials in their relationship to processes of knowledge generation in the interaction between human and artefactually contained knowledge. A broad framework can, subsequently, incorporate solutions to particular issues or tasks, although always within the underlying interpretive framework and not in isolation. These past documentation projects can inform present and future academic practice in documenting and handling language data (Thieberger & Jacobson, 2010), feeding back into the theorisation of meta-documentation as an inductive approach (Austin, 2013). As illustrated in Figure 2.1, legacy materials can create a channel of communication between present and past generations of speakers, researchers, or archivists, and we rely on these past generations' knowledge to bridge gaps in our own understanding of the artefactual legacy. At the same time, 'interpretive challenges are inherent in the nature of all records produced by others

at another time for their own purposes' (Dobrin & Schwartz, 2021, 5). This turns the work with legacy materials into a task of interpretation and reconciliation between different forms of knowledge. These tasks are difficult, if not impossible, when contextual clues are missing (Latour, 1987); if we consider publication and reuse of data as a recontextualisation that simultaneously implies a dissociation from original contexts, the framework of linguistic legacy materials must cover instances of reuse, collation, or transfer into other analytical or spatiotemporal contexts, e.g. data citation tracking (Paper IV) and misinterpretations of linguistic examples in the literature (Engh, 2006). As these questions of data integrity and reproducibility are topical in linguistics, a theorisation of linguistic legacy materials becomes increasingly significant.

Chapter 3

A framework for linguistic legacy materials

This chapter draws from experiences working with the Kraasna legacy materials to construct a dynamic framework for processes of interpretation and meaning-making. The categories and domains it introduces are to be understood in a broad sense that require us to define and delimit them for each context of working with legacy materials. This approach to the construction of the framework is intended to be mindful of ambiguities or intersections in the nature of legacy materials that cannot be mapped or categorised in a strict sense; any user of the framework needs to consider the interpretive work and positionalities they bring into the picture by applying this theory. My work is thus claiming its heritage in the humanities and its philological and hermeneutic traditions which stand against approaches in many natural sciences that claim to create 'unambiguous, comprehensive, and correct representation of at least some domain of reality given the caveat that our knowledge is always limited and human errors are always present' (Wendell Compton, 2014, 431). While accepting the premise of limited human knowledge and fallibilism, I contend that ambiguity and representational choices are inherent to the meaning-making process, where the human agent cannot be separated from their description of reality.

The framework in Table 3.1 maps the concepts *level of reference* on the *object of study*. The object of study may be abstract or material, i.e. knowledge held by human memory, or an artefactual representation of it. While the initial model in Table 1.1 contained a meso-level, this framework operates with a binary distinction of the polar micro- and macro-levels, which rest on a continuum where a meso-level is conceptually possible but not part of the basic form of the framework. This binary distinction allows for the definition of the micro- and macro-levels as treating the object of study either *as* a system or *in* a system, i.e. looking at internal features of the object or in relation to external knowledge and artefacts. Importantly, this framework is to be understood as dynamic, where the individual domains, i.e. the cells of the table, are less important than the trajectories and continua that span the dimensions of the table. While description can take place within a single domain, knowledge creation involves a transfer between these planes. This is where the human role is crucial, as we are moving and translating between planes, thereby making meaning in

the transfer of knowledge. Inspired by the concept of the 'social lives' of linguistic legacy materials (Dobrin & Schwartz, 2021), I call these human-centred movements of knowledge *social dimensions*. These lie between artefactually contained and abstract knowledge, as well as in the abstraction from microscopic to macroscopic levels and vice versa.

		object of study	
		abstract *focus on knowledge* *humans*	**material** *focus on artefacts* *non-humans*
level of reference	**micro-level** *internal* ***as*** a system	(Paper I)	(Paper II)
	macro-level *external* ***in*** a system	(Paper III)	(Paper IV)

Table 3.1: Domains of knowledge creation

Importantly, although my work is focused on researchers and institutionalised processes of knowledge generation, the social dimensions are not limited to research. Any movement between these planes can be captured by the framework, e.g. if a member of the speaking community searches for legacy materials left by a relative, where the transfer of knowledge and information about the real person is checked against the metadata in the archive. If a relative can identify the speaker in the artefacts and link them to memories of the existing person, the user of the archive will know more about the consultant than is contained in the material domain, therefore knowledge is generated. As a consequence, the understanding of artefacts in the material plane is broad. It contains any medially stored information in texts, in the broad sense relating to knowledge (as applied in textual criticism, e.g. Derrida, 1997), implying communication across an 'expanded speech situation' where the interlocutors are not in the same spatiotemporal context ('zerdehnte Sprechsituation' Ehlich, 2007). Likewise, (multi-)medial inscriptions, physical objects, 'paratexts' (Genette, 1997) and non-human actors can be mentioned here, with which we can occasionally interact as if they were human (Latour, 1993 [1991]). The latter becomes especially relevant in the context of knowledge and memory institutions which crucially shape possible interactions with legacy materials. Yet, as mentioned above, our experiences based on reflective practice help us to establish which humans and non-humans are relevant in the contexts we are describing (Van Manen, 2014).

The framework is influenced by social constructivist understandings of knowledge arising from interaction with others. This can be seen by the fact that humans are both situated in the definition of the abstract plane as well as responsible for the transfer of knowledge between planes. For this reason, documentary linguists are not only recording information on the consultant they worked with but also note information about themselves and any other humans who were present during the interviews, in an attempt to keep the human

influence visible on all levels. This focus on the 'human in the loop' (Bird, 2020) also affects computational movements, where human motivation is not directly visible or causally linked to an individual transfer between planes, yet affects the underlying understanding and motivation for adding value or recording knowledge. Thus, any automatic tagging, metadata completion, compilation or collation by a computer requires a documentation of the human hand in the creation of the script or algorithms. This is due to the intermediary role taken by technology in these cases, which conducts the interpretive movements between planes. The researcher does not have to interact with the object of study directly, yet communicates their motivations and understandings through the way of designing the tools. This also affects machine learning algorithms that are trained on data sets rather than receiving direct instructions from the creator; there are motivations in the choice of training data and the decision to employ the application for a particular case (i.e. the user deciding that the results yielded by the algorithm appear plausible). The movements between planes do not have to be value-adding in order to constitute knowledge transfer or creation[1]; even in using media as bearer of knowledge, e.g. by reading them, we are translating the inscribed information based in signs and symbols into meaning. By the same token, the creator of a computational tool defines the parameters of the output they wish to receive as in itself interpretable, e.g. text or numbers. In each of these *uses* and *reuses* (Kenfield et al., 2022), there is an inherent human perspective to the interpretation and movements between planes which is not exclusive to linguistic legacy materials but any artefacts that are designed as bearers of knowledge.

3.1 Artefacts

As discussed, artefacts in the framework are to be understood in a broad view including not just media containing language data – like manuscripts, audio and video recordings, or digital corpora and data sets – but also descriptions and 'paratexts' that help with their interpretation. In this respect, field diaries or scholarly communication such as correspondences or publications can be included in the framework, as can be analyses or discussions that add context to the data (see section 2.1.1 for examples from the Kraasna case study). Overall, an artefact is an object or medium that contains a snapshot of human memory and contextual knowledge at the point of recording. It does not need to be conscious knowledge but includes notions that the creator is subconsciously aware of, e.g. technical specifications that are part of the metadata but implicitly known to the documenter who

[1]Knowledge creation, in this understanding, is not tied to originality but the processes of meaning-making. This also affects those natural sciences that claim an absoluteness of their theories, e.g. mathematical logic and arithmetic: While there is no doubt that the sum of one and one is two, there is a human movement required in the interpretation of the notation '1+1', where the symbol + indicates the addition of two numbers (the same symbol is also used in some programming languages for the concatenation of two strings, i.e. '1'+'2' yields '12' as a sequence of symbols). At the same time, the mathematical notation requires the application of knowledge in the form of rendering sums appropriately – if we are using a binary numeral system, the correct notation yields $1 + 1 = 10$ where $(2)_{10} = (10)_2$. The related issue of conventions and standards will be addressed in sections 3.2 and 3.3.

decided on the use of particular technological tools. The more these metadata are tied to the material domain, the easier it is to reconstruct them from the artefact itself, e.g. brand names of technical equipment; decisions about data formats or settings are less observable and border on the abstract plane from where the researcher draws the justification for their decisions. The act of inscription that transfers memory onto a medium is thus containing knowledge and perception in context, where contexts also involve subconscious, ephemeral notions. For ethnographic writing, Clifford (1990) outlines three processes of recording: inscription, transcription, and description. These are differentiated by their embeddedness into the context of the observed situation, i.e. the interview, and the ethnographer's vantage point in creating the record. These three process fall onto a continuum of increasing distance from the inscribed event to its description afterwards. The creation of any artefact is therefore not just a matter of inscribing reality but is a socially constructed and human-mediated 'exscription' (Nancy, 1993) of a 'thought of the sense of the world' (Nancy, 1997, 9) that in itself is interpretation beyond the inscribed form. Anything that is documented is thus exscribed and contains a social dimension of human meaning-making. When working with these artefacts of documentation, we need to move beyond the physical nature of the legacy materials and uncover their relationships to ecologies of knowledge within and outwith. This holistic understanding of linguistic (legacy) materials as results of documentary activities goes beyond the 'ecology of documentary and descriptive linguistics' offered by Good (2007), as it includes knowledge and artefacts in a broad sense and is not just confined to the academically oriented endeavour of language documentation and description.

At the same time, for linguistic legacy materials, the focus will initially rest on the material plane, as any embedded knowledge needs to be reconstructed from the artefacts at hand. Paper I is a study of the Kraasna documentation projects as a closed system, yet looking beyond the artefacts to investigate the contexts of creating these materials. This follows the idea that documentation in itself is a communicative act, where every little mark in the materials can matter – not just in linguistic annotation but in the overall creation of a written document. A symbol may add emphasis, highlight important information, guide the reader, explain content or add context, flag issues, or may simply be doodles or smears (Marshall, 1997). The link between documentation and communication is illustrated in the definition of a document offered by Windfeld Lund (2010, 744)

> 'any results of human efforts to tell, instruct, demonstrate, teach or produce a play, in short to document, by using some means in some ways, is very focused on activities around making documents, in other words on practices.'

Therefore, we can assume that documentary artefacts are intended to communicate knowledge between the context of creation and the subsequent use in transcription or description. This communication can be either oriented towards the creators themselves, e.g. Kallas' fieldnotes, or a wider audience. The researcher's agenda and communicative intent can be seen in the narrative created by their project, as the artefacts bear traces of their creators (cf. Halbwachs, 1980; Appadurai, 1986). At the same time, the work of editing and curating legacy materials also folds us as secondary researchers into the picture we are trying

to reconstruct. Following Gorichanaz (2019), any document has not only intrinsic and extrinsic (i.e. physical and attributed) information but also involves abtrinsic and adtrinsic information, consisting of feelings and memories that are ascribed to the artefacts. The ab- and adtrinsic moments are highly relevant in the movement between material and abstract planes – what is the reader thinking and feeling when making meaning of an artefact? In this process, there is not only the positionality of the original researcher but also our own when we reconstruct and reinterpret what has been deposited in the past. A thorough account of meta-documentation, i.e. 'documentation of the documentation research itself' (Austin, 2013, 6), can help in the reconstruction but will not offer a neutral perspective, just as any reconstruction will not be neutral in itself. In this view, we are still adding to the meta-documentation and the narrative surrounding an artefact, even outside of the context of the original research project (Nathan, 2010).

Since many issues regarding linguistic legacy materials occur at the interface between textuality and practice, I initially opted for philological methods in the reconstruction of the meta-documentation, including critical edition and interpretation of the textual artefacts. Philology, in this view, covers all activities that aim to 'preserve, monitor, investigate, and augment our cultural inheritance' (McGann, 2013, 334) which render it a 'fundamental science of human memory' (ibid., 345). This link to human memory can be established when conceptualising the work of the critical edition of legacy materials not as hunting for divergent versions and 'reading [other researchers] against their data' (Dobrin, 2021, 39) but as the reconciliation of the present-day researcher's knowledge with the knowledge contained in the brief snapshot of human memory in the artefact. This knowledge is only accessible through interpretation of the materials and, in turn, contextualisation amidst other knowledge or materials, or, as Assmann (2008, 97) describes it, 'recalling, iterating, reading, commenting, criticizing, and discussing what was deposited in the remote or recent past'. My decision to use philological methods was likely facilitated by the fact that the curation would stay within the same tradition that created the materials, as Kallas and Ojansuu were both philologically oriented with respective interests in folkloristic and linguistic research. Yet, on a theoretical level, the 'critical edition of legacy materials' (Linguistic Society of America, 2010) involves hermeneutic and philosophical questions that go beyond traditional philology. In practice, the textual nature of the linguistic legacy materials and the type of data that may be contained within make philological methods and tools useful. Depending on the context, however, the inclusion of other disciplines is important, particularly ethnography, arts, technology and information science, library science. In the work of curating linguistic legacy materials, the curator needs to decide which materials, paratexts, and real-world knowledge are relevant in the reconstruction and interpretation. They need to adopt a reflexive and interdisciplinary stance that can help to uncover relationships between knowledge and artefacts and the multiple contexts of creation and reception thereof.

In this view, the use of the term 'philology' is not identical to the (Classical) philology as a study of historical texts or comparative historical linguistics. What unites the philological approach to linguistic legacy materials and the traditional perspectives is the desire to understand more about a textual artefact, its contextual embedding, and, through the

analysis, humankind itself. The same way Bronze Age writings need to be deciphered, a philologist working with legacy materials needs to apply a range of skills to make sense of the inscriptions by inference, comparison, or forensic analysis. Contextual knowledge about the medium or the processes of writing associated with the artefact are likewise important, adding a material domain to the object of study. At the same time, understanding a textual artefact among other texts (externally) and within the social fields where knowledge circulates, requires the investigation of society and its positions towards the artefact and its contained meaning. Like a classical philologist aims to reconstruct information about the past and gain insights to historical societies, the philology of legacy materials reconstructs and contextualises the circumstances in which knowledge was generated, documented, and communicated. In this process, philology draws from different disciplines that have, traditionally, been closely linked to the study of historical texts and language itself. Yet, the call for a philological stance is not just going back to the roots of this discipline but calling for a modern 'multidisciplinary' combination of skills; bringing back the holistic perspective of humankind and its relation to textual artefacts. Additionally, new disciplines that have historically been outside of the philological canon can be involved in the process, for example indigenous studies. The goal of this combination lies in the preservation of a multiplicity of perspectives for the co-construction of meaning. While this was not the objective of traditional philological research, the curiosity about a historical perspective on the artefacts and with it a different understanding of society, humankind, and knowledge is also ingrained in this discipline – philology has always sought out new insights through the eyes of authors in the past (Foucault, 2010 [1969]).

If we consider Kallas' data as an example, the recomposition and editing of language examples makes him as much a part of his published data as his consultants. While the additional linguistic filters he added during the transcription reinforce the notion of what could be called "Kallas' version" of Kraasna, even the inscription itself is not neutral. As Clifford (1990, 57) reminds us, 'noting of an event presupposes prior inscription'; there is implied knowledge in the artefact that we need to reconstruct to understand the contexts (Thibodeau, 2001; Huvila, 2015). At the same time, this prior inscription also affects us as readers and curators, when we are reconciling the artefactually mediated with our own knowledge. For example, remarking that Kallas held a nationalist stance that could have affected the way he describes and depicts the language and its speakers. This is my critical reading of his artefacts, including his 1903 monograph as a paratext, which links myself to the interpretation; a curator or editor cannot remain anonymous in the process, as their stance is neither objective nor neutral, as if standing outside of the system. Thus, 'without careful analysis of *why we find what we do* in all the sources, we may draw unwarranted conclusions about the language' (Paper I, 91), just as readers of our own reconstructions may. Paper I therefore concludes that 'even the thickest metadata cannot replace philological care', as neither data nor metadata can speak for themselves. This is also echoed by other scholars working with linguistic legacy materials, e.g. Dobrin & Schwartz (2021, 19) highlighting that '[a]s with metadata, pragmatic annotation can only capture a small fraction of what a researcher prospectively imagines an unspecific future user would want to know'. The assumption that a sufficiently large amount of data or

metadata could replace the human work of contextualisation is seen as a Western bias by Tuhiwai Smith (1999), based on the dominant materialistic paradigms in institutionalised Western knowledge systems (cf. Wendell Compton, 2014).

'Thick' metadata (Nathan & Austin, 2004), as a desired feature of the meta-documentation in present-day documentary projects, needs to be understood as an interpretive layer in the artefacts. It involves a movement between planes when the researcher records or ascribes information that is otherwise not contained in the material domain to the artefact. This additional information can take various shapes, including ab- and adtrinsic information like the reader's reactions to the material (e.g. Schwartz, 2021), and stem from a variety of sources[2]. The addition of the appendix in ES MT 224 mentioned in chapter 2.1.1 can be seen as a value-adding process, whereby the author combines at least two sources of information into a single artefact. This combination of different sources of knowledge also informed the methodology for the linguistic description of the recordings, assuming that Ojansuu's manuscripts bear traces of his cognitive processes. Even if they are only contained in his 'headnotes' (Sanjek, 1990), the interpretation of the recordings alongside the manuscripts in the processes of transcription and description helps to base the transcriptions not only on my impressions of the, at times barely discernible, speech but also the impressions of a colleague present during the original event. This does not mean that any decision by Ojansuu would be accepted without critical questioning, but in adopting his versions, I need to take responsibility for giving credence to the decisions of his I adopted. However, the negotiation between artefactually mediated information and Ojansuu's and my interpretation as human filters offers an alternative to a mechanistic interpretation that would take artefacts at face value (Bird, 2020). While boundaries between the various roles are blurring (Weber, 2021c, Paper III), this can be mitigated through an open and transparent documentation and communication of decisions.

In times when publications and data sets are seen as validation of one's membership in the academic community (Bond, 1990), the role of recontextualisation of linguistic legacy materials needs to be investigated alongside other academic activities. Paper III addresses the need for preparing students, as well as established researchers, for the tasks that stem from 'curation' as the effort of reconciling different sources of knowledge and learning more about past practices that influenced the artefacts we find in the archives today. At the same time, this involves a critical reading of past publications, which function as a 'commentary' or 'paratext' to the artefacts (Petőfi, 1973; Genette, 1997). In doing so, we renegotiate meaning with the previous generations of scholars and consultants, provided that contributions are credited appropriately and any changes to the materials are attributed to an author, ideally with 'thick' metadata about the underlying motivation (Nathan & Austin, 2004; Austin, 2017; Thessen et al., 2019; Andreassen et al., 2019; Bird, 2020). This goal of keeping full records of human agents goes against the trend of 'purification' that is attempted in many natural sciences (Latour, 1993 [1991]), and increasingly adopted in the

[2]The framework in Table 3.1 assigns the papers of this dissertation to the domains in which their main viewpoint is located. Yet, they all draw from different planes to create knowledge rather than listing observations or general knowledge. In order to understand all factors in the reconstruction and to analyse my own positionality, it is necessary to read and interpret these publications in context with each other.

social sciences and humanities. The alternative of tracking all human interaction, including accessing, downloading, or reinterpreting might appear more complex than attempting to remove the human influence from the records altogether. Yet, it enables a later critique and reconstruction that is impossible if information or contextual clues are incomplete. The social lives of the artefacts evolve when we, likely as part of our everyday practices as researchers, interact with artefacts and data, which renders them our own versions of them. While some authors have used the criteria of changed ownership and originality to differentiate between 'use' and 'reuse' of data (Pasquetto et al., 2017, 2019), even an annotated copy of a journal paper or a local copy of data on my hard drive can contain added value beyond the archived or published versions. This can help curators with the reconstruction of research trajectories and processes of meaning-making.

My outline of curation goes beyond what other authors have considered to fall under this area of academic activity. Muñoz (2013) describes it as the 'development of indices, annotated linguistic corpora and digitally encoded texts', while Xie et al. (2022) emphasise the active selection and arrangement by experts and information systems, continued care for content, as well as the collection, classification, and provision of information to a user of a curated collection. While these authors show that curation is not just a matter of preservation and presentation of archival records, Higgins (2018, 1326) marks that the conscious design of a collection has a community of users in mind and aims to create trust through human judgement. This ties back to community relations and involvement (see Figure 2.1) and introduces the crucial interpersonal factor of trust. Beyond trust, academics should aim to show their appreciation of speakers' efforts in the past and present who participate in language documentation projects – a different social dimension to the artefacts produced through research activity. As several position statements outline, it is important that merit is given to all contributors whether they are academics or members of the public (Andreassen et al., 2019; Berez-Kroeker et al., 2018; Linguistic Society of America, 1994, 2010, 2018), which can be achieved through the mediation of resources and thorough meta-documentations. As a positive effect of mediation, Kurtz (2010) found that it helped to generate more complete deposits and accurate metadata (on the concept of mediation in linguistics, see Holton, 2014). Although this technical focus is not as important in the scope of this dissertation as the mediation of research narratives, it can be seen as a welcome effect of engaging with archival deposits. With a focus on mediation, I define the function of curation 'as an external locus of reflection and negotiation between the communities of documentary and descriptive linguists [which] reviews procedures and outcomes, checking for accuracy or "replicating" results, and "mediates" between the needs of various groups' (Paper III). This requires a self-critical stance in knowledge generation (Topp, 2000; Gourlay, 2006; Tsoukas, 2009), as well as archival infrastructures that support curation.

To conclude this section, I will illustrate the example of anthroponyms on the different social dimensions in language documentation artefacts (Weber, 2021b). Kallas and Voolaine offer information and metadata on their consultants in their fieldnotes, while Ojansuu does not. Ojansuu's manuscripts only offer first names and in one instance a patronym. In this interpretive step, the link between the material and the abstract do-

mains becomes visible – if we assume that all names refer to existing people, we are using their givenness and referentiality that permeates the material. While their referential function is without major problems on the internal domains, i.e. references within a text or consultants within a documentation project, the transfer to the external planes becomes difficult. Anthroponyms are determined by a third party and can be subject to symbolic power (the adoption of Russian naming customs and conventions by the Kraasna community in the early 20th century might have been an attempt at preventing discrimination), yet they are not globally applicable or intelligible. Even within the Kraasna documentation, we need to ask whether consultant *Ulĺa* in Ojansuu's manuscripts is the same *Ulĺa* interviewed by Kallas. Certainly, consultants will already have names at the start of any documentation project, yet these can also change or be adapted to resolve issues with metadata or project management. It is, thus, important to record variants of names throughout the project in a hope that the referential contexts remain at least reconstructible in the future, i.e. future generations can infer who the mentioned individuals were, especially in relation to other community members. In this process of curation, I also pondered whether the anthroponyms at the top of Ojansuu's manuscript pages were pen names to anonymise his consultants, in case the data was misused in politically unstable times. The problem would not exist within the internal domains, i.e. references within a text or linking materials to speakers by a pseudonym is not problematic, as long as the movement of this knowledge between planes is not attempted. As soon as we try to link referents in the text to pseudonyms, ask the community about these people, or query archives for more artefacts by these consultants (an issue encountered by community members looking for recordings of their relatives, cf. Khait et al., 2022), we would need access to Ojansuu's 'headnotes' to resolve these issues. Although I do not expect that the names are pseudonyms but rather minimalistic accounts of metadata, the difficulty in interpreting and linking them between projects stems from the lack of contextual information. Only Ojansuu could provide the necessary information to transfer knowledge between the planes, but this should not discourage us from interpreting the narratives as ethnographic resources.

3.2 Knowledge

In the social constructivist view of knowledge creation presented above, fieldwork and documentation constitute communicative acts. Consequently, we must always interpret our own data as a result of communication, i.e. it cannot be neutral, involving different layers of positionalities that are embedded into the records (Dobrin, 2012; Moore, 2009, 2013; Nevins, 2013, 2015; Schwartz, 2021). For example, consultants make decisions about what they tell about themselves and their community in a negotiation between their own social position within their group and their self-understanding as an interview partner in a documentation project. This also involves a careful consideration about the role of the fieldworker, as well as any bystanders or envisioned communities accessing the materials afterwards. The researcher likewise approaches the fieldwork encounter and each interview with an idea or expectation in mind, at least about the objective of the documentation

project – this affects the questions asked, the positions expressed towards the speakers and their language (e.g. Kallas' "colonialist" style), the form of interview (cf. Beer, 2021), as well as decisions about transcribing and analysing the language (see for example Paper I). These positions of the researcher influence the consultants and assistants, either through direct instruction or indirectly in the act of documenting. In the case presented by Dobrin (2021), a transcriber had taken the agency to 'correct' the recorded language for the transcription, adding a new layer of interpretation to the artefact, which enables us to investigate positions and understandings of the speakers, assistants, and the researchers who worked with these data. As interpretation is a movement between different domains or planes of knowledge, we can observe the social dimension of linguistic legacy, as '[t]hese acts of interpretation – even problematic or "incorrect" ones – can generate insight into both linguistic structure and the social relations developed and mobilized in the context of research' (Dobrin, 2021, 37). The link between documentation and communication is also observed outside of documentary linguistics[3], where *documenting* implies an intentional act that aims to highlight or communicate a particular action, interaction, or practice, even if it is tied to a physical object that is archived. Without this purpose behind the documentation, the record or 'collection' – which already implies the question *of what* – is meaningless if we do not reconstruct a context and links to practices or other artefacts, i.e. generating knowledge ourselves, as is required in the work with linguistic legacy materials.

In the linguistic literature, an increased emphasis is put onto 'reproducibility' as an ideal that linguistic research should aspire to reach (e.g. Berez-Kroeker et al., 2018). This goal is borrowed from natural sciences and implies a mechanistic process of knowledge generation that takes data and 'produces' an analysis of them. In contrast, the above mentioned interpretive work of meaning-making in context would ask for linguistic research to be retraceable and the underlying contexts of documentation and interpretation reconstructable. Therefore, the entire linguistic enterprise, with a few exceptions (Weber & Klee, 2020), should strive for 'reconstructability' rather than 'reproducibility'. The process of knowledge generation is fundamentally social and the interpretations and conclusions about language as our object of study are formed as an 'intersubjective objectivity' (Longino, 1990, 2002). This concept describes a method of reaching a consensus under the premise that there is an open exchange of ideas, criticism, and innovations to which the community and its knowledge system responds openly, and if participants can be assumed to have an equal capacity to make judgements under publicly communicated standards of knowledge generation. That means, an open discourse about interpretations of language and criticism of theories and methods can lead to collective meaning-making among scholars and stakeholders which subscribe to the standards and assumptions of the academic discipline. An understanding of knowledge generated as 'intersubjective objectivity' emphasises the social element of communication, where 'objective knowledge is

[3]Windfeld Lund describes the link between documentation and communication as a complementary one. However, we see overlaps in the area of language documentation, where the object of documentation is in itself communication: 'While communication is biased towards the issue of sharing something among a group of people by the prefix com-, documentation may be considered to be biased towards the very act of using some means in a certain way' (2010, 744).

asocial knowledge' (Dobrin & Schwartz, 2021, 9), in that it is not contextualised in any meaningful way to the human reader meaningful way. In respect to the investigation of artefacts of past and present language documentation, we are simultaneously exploring the authors' own interpretations and reconciling their knowledge with our own in the hope of reaching a consensus. It is a social process that transcends the temporal confines of the original documentation or description, whereby we interact with past generations through the textual artefact as the medium for the extended speech situation (Ehlich, 2007). The focus on individual narratives and autoethnographies in our research can evoke criticism of a perceived 'anarchy' (Bloor, 1991), if conventions and consensus do not mediate between our individual, subjective interpretations (Kuhn, 2012 [1962]). This might create the impression that robust standards could solve any issues pertaining to reproducibility. Although standards are useful and can mitigate some problems, these standards must not be understood as merely technological but also social conventions, since 'language data is a social product' (Beer, 2021, 133). This means that standards are important in facilitating interpretation and interaction with data but also act as a restricting factor to who can access and interpret data and in which ways this is possible (cf. Weber, 2021c).

Accessing and interpreting artefacts, i.e. interacting and reusing them, implies that a successful knowledge transfer has taken place. This, in turn, suggests an instance of communication (Markus, 2001) – in the most basic sense – as a combination of intrinsic, extrinsic, abtrinsic, and adtrinsic information (Gorichanaz, 2019) flowing between the artefact, its creators, and its observer, i.e. a sign and its interpreters in a semiotic understanding. Through interacting with artefacts, we are interacting and positioning ourselves in respect to others and a general public. By reusing artefacts of language documentation and language data, we are entering a discourse with other researchers, including those who created or deposited a data set. The work with linguistic legacy materials and other artefacts of research is an inherently social process (Yoon, 2017). In order to participate in this process, one must be familiar with the standards and conventions in the field. Khait et al. (2022) describe that community engagement can be hindered if this participation requires literacies, e.g. computer, data, or even basic literacy, which the target audience does not possess. In this understanding, standardisation is not just a matter of agreeing on rules or conventions for a small academic audience, but it denotes minimal requirements for any user of the artefacts following the standard. Users of language data, and not least 'reusers' such as linguists or curators, need to be taught how the standards shape our interpretation and interaction, e.g. transcription systems or file formats, while also learning to reflect upon the standards themselves (Paper III). In terms of a specific data literacy, for example to work with language data, Calzada Prado & Marzal (2013) mention the abilities to access, interpret, critically assess, manage, handle, and use data ethically, e.g. in their preservation and curation. While basics of data literacy can be expected from any researcher, particular issues in data management or the outline of ethical guidelines governing the use and reuse of sensitive language data need to involve experts from different communities. Basic knowledge of a tape recorder's functions does not guarantee that an untrained user will always handle a recording with the required care that protects the medium from any damages. Likewise, modern standards and guidelines for ethical field-

work cannot replace the respective language communities' insight into appropriate and responsible handling and provision of the data by researchers and archives (e.g. Dwyer, 2006; Austin, 2010). It is essential that the various communities come together in the documentation, curation, and description of language data and the media containing them – the social process involves participation of all stakeholders. Thus, meaning-making has not only an internal social dimension but also relates different communities and domains of knowledge.

The involvement of different stakeholders affects data quality on different levels. Koltay (2015) delivers a compelling account of the factors at play: Trust in a data set depends on its authenticity, acceptability, and applicability, and arises from the creator or curator's relation to the user. Authenticity, in this definition, reflects a measure of proper scientific methodology, i.e. instruments, frameworks, completeness of data, accuracy, and validity; in order to assess authenticity, data need to be understandable and interpretable by the user. In order to facilitate data use, data sets must be discoverable, accessible, and interoperable (i.e. FAIR, Wilkinson et al., 2016), enabling further validation and discussion of results. Although this definition emphasises technical factors, these are just conducive to the communication and social construction of meaning. If instruments or methods are not recorded and communicated appropriately, any secondary researcher may not be able to understand and interpret the data yielded by a research project, precluding any discourse on the results. There appears to be an interplay between these technical and conventional aspects of standardisation and the possible interpretations and interactions they facilitate. At the same time, 'linguistic legacy data can never be so self-explanatory that those involved in its creation can be disregarded, or the task of reconstruction is rendered unnecessary' (Dobrin & Schwartz, 2021, 23) – the human factor deciding on and implementing standards in any given documentation project needs to be recorded, as it contains the key to the interpretation and assessment of an artefact. In the evaluation of data quality in the above understanding, checking the potential reconstructablity of contexts is a routine task that should not only become relevant after a 'disaster' (Meister Ko. Freitag, 2022) has happened and knowledge has been lost. The prevention of this knowledge loss is a proactive process which begins at, or even predates, the time of documentation as the inscription of a snapshot of human memory into an artefact. In the interpretation and reconstruction of contexts we are constrained by what has been documented, without any contextual clues we cannot reconstruct or infer knowledge from other domains, e.g. the history of a field, scientific beliefs, conventions or 'best practices'. At the same time, working with linguistic legacy materials involves a renegotiation of knowledge and methods used in language documentation. Consequently, it opens avenues to explore and correct historical injustice, e.g. nonchalant views of speakers in colonial settings of language documentation, and allows for community involvement. Members of a linguistic community may reassess legacy materials for their appropriateness and can reinterpret them into a version that they want to display or communicate about themselves and their language (Schwartz, 2021).

In the negotiation and social construction of meaning, different perspectives on abstract and material knowledge need to be reconciled. In other terms, phenomenological observations and epistemological explanations about artefacts and their link to 'reality' are

discussed, where 'understanding' is inherent in both (Gorichanaz, 2017). In a Heideggerian reading, epistemology and ontology are identical, in that our impression of reality will always be a human-based *aletheia* – the social dimension of knowledge sets the stage for discourse and the co-construction of meaning. These views 'help account for phenomenological and critical analysis of understanding of reality with respect to the digital' as well as any artefactual or medially contained representation of reality (Wendell Compton, 2014, 439). It is, thus, important for the humanities to establish approaches to the scientific enterprise that consider social and human, i.e. phenomenological, perspectives in a broad sense without establishing an authoritative version of reality that emanates from our data (Seidel, 2016), as attempted by 'reproducible' sciences. In contrast, the movement between planes in the framework of linguistic legacy materials adds meaning to what is perceivable, thereby combining ontological and epistemological approaches. The interaction with the materials and contexts in which they were created and reused constitutes a hermeneutic process (cf. Hjørland, 2005), which emphasises the human role in knowledge creation.

3.3 Archives

In the discussion, it is important to consider means of accessing and interacting with artefacts and data, as a way of creating a forum for negotiation and generation of knowledge. The archive is an institution and location where both access and curation takes place, as it holds artefacts and facilitates access to them. Yet, in doing so, it does not take a neutral position between depositors and users, communities and researchers but is adding a layer of interpretation and positionality. As Derrida (1995, 12) describes '[a]n eco-nomic archive in this double sense: it keeps, it puts in reserve, it saves, but in an unnatural fashion, that is to say in making the law (nomos) or in making people respect the law'. By the virtue of deciding on standards for categorisation and presentation within the archive, it actively filters and adds meaning to what is preserved. While the categorisation by provenance – the 'principle of provenance' (Abraham, 1991; Douglas, 2010; Ross, 2012) – is most frequently applied as a seemingly neutral approach to managing artefacts, a thematic collection requires curation and interpretation by the archivist or curator. Yet, even in the process of adding descriptive metadata or information to an item, the archivist employs their individual perception of the artefact or commonly agreed standards within the community. This ties back to the practice ascribed to an artefact and the understanding of its purpose, meaning, or relationship to actions, as well as other artefacts. Tuhiwai Smith (1999) reminds us that the Western perspective employed by many knowledge and memory institutions can disenfranchise indigenous knowledge systems, since archival records and deposits are described from this external position and not within the knowledge system to which they are native. This already affects the initial evaluation and judgement of whether or not an artefact or a data collection should be archived and to what extent it is to be preserved. Since physical and digital storage is limited, the archives, as organisations, need to manage their capacities by deciding what they want to acquire or need to reject – this evaluation does not make rejected artefacts or legacy materials less valuable (Beer, 2021).

As argued in Paper III, any interpretation or curation has a value that lies in the addition and negotiation of knowledge between the user and the artefact, rendering these processes worthwhile and valuable in themselves.

In Paper IV, I consider the language archive as the 'central locus of academic activity, where present-day research and historical research are reconciled', emphasising the role of linguistic legacy materials within modern research infrastructures. At the same time, the focus on research should be extended to cover all processes of knowledge creation and meaning-making, especially those by citizen scientists and community members. As discussed in the literature, archives need to be designed in ways that facilitate discourse and active participation by the public (Huvila, 2008; Woodbury, 2014; Henke & Berez-Kroeker, 2016; Wasson et al., 2016; Wasson, 2021). If the process of curation involves different communities and stakeholders, the participation in meaning-making through archives and their deposits can enrich both the academic and the community life. It reconnects generations and establishes communication between these groups, e.g. if a community aims to teach historical language practices and asks a scholar for their support. Yet, the archive should also allow for indigenous knowledge to flow between historical and present-day communities without the mediation of an archivist or linguist (cf. Duarte & Belarde-Lewis, 2015). In order to keep these artefacts and the contained knowledge visible within discourse and processes of meaning-making, the fixation during the points of creating and archiving needs to be dissolved, e.g. by constantly reviewing and updating materials and information (Babinski et al., 2022). This can happen through human curation or with the help of computational tools, which track the history of interactions with an item after it has been archived. At the same time, this process requires more than just updated metadata or a download tracker; the actual use and reuse of data, as interactions with artefacts and items, involves the combination of metrics with interpretive work which includes discussions or instances of knowledge generation that take place outside of the archive. For example, a link to paratexts like (scholarly) publications can enrich the archival record and supply context to it. Given that there are frequently differences or re-interpretations between the original data in an archive or repository and the published data (Engh, 2006), these must be critically examined to understand how an artefact was used to generate new insights. In this view, scholarly communication like publications turn into a new primary data source that could be archived or at least contextualised alongside the legacy materials (Haendel et al., 2012; Weber, 2020b), i.e. in the external material domain.

In the theorisation of language archives, the focus often lies on technical details (Wu & Chen, 2022; Bharti & Singh, 2022) or considerations about access and rights (Austin, 2010; Nathan, 2010; Woodbury, 2014; Nathan, 2014; Yi et al., 2022). In these accounts, metadata and descriptions are often based on standards that allow for some control over the archival deposits, which are, of course, necessary in the creation of an archive. At the same time, the dynamic language archive that facilitates exchange and knowledge generation needs to go beyond instructions and conventions for depositing and using artefacts. These formulate the broad goals to which all archives should subscribe, such as the need for findable, accessible, interoperable, and reusable data (Wilkinson et al., 2016) and the aspiration to generate collective benefit, giving authority and control over the data to the communities,

in order to document and archive in a responsible and ethical manner (Carroll et al., 2020). In addition to FAIR and CARE principles for indigenous language data, meaningful interaction and sensible negotiation of meaning are required; these are supported by FAIR and CARE standards, but do not naturally follow from them. Scholars and archivists need to accept the responsibility for the application of standards and not hide their own agency behind these abstract conventions and rules. These are just the basic requirements for the subsequent discourse and negotiations, in which the human role shapes any interpretation of the artefact, its meaning and value, or relevant processes in archiving and displaying it. During these processes, indigenous knowledge and perspectives from the communities are highly relevant in defining access, use, and reuse. All decisions that are made in respect to the artefact need to be recorded without supplanting metadata that is already in place (Andreassen et al., 2019). An edited transcription of language data like in Paper II bears traces of the initial transcriber (in this case Ojansuu) and the curator (i.e. myself) – although I accept full responsibility for "my" transcriptions, it is necessary that the underlying interpretations by Ojansuu are not lost from the metadata. The newly created data set, as an artefact in the research process, is a result of the negotiation and social co-construction of meaning between myself and Ojansuu, indirectly facilitated through his legacy materials. In a conference paper (Weber, 2020a), I considered means to keep changes attributable to each curator under the idea of 'metadata inheritance' (Greenberg, 2009; Niu, 2013), such as a log of changes allowing for the reconstruction of versions at different stages within the curation.

The ultimate goal consists of transparent accounts of interactions with an archival record, which allows us to reconstruct – either manually as in the Kraasna case or through computational means – processes of interpretation on a given version of an artefact in particular contexts of knowledge generation. Therefore, "the digital language archive is not just the beginning (e.g. data discovery) or the end of the research process (i.e. data deposit), but takes a central role throughout the entire research trajectory, beyond the end of any project. In this view, it mediates between individual projects, different artefact types and various stakeholders" (Paper IV). Following this view, modern archival infrastructures can support the reconstruction and foster transparency of past and present research by dissolving the strict boundaries of archiving as fixed to a certain point in time. As discussed by Babinski et al. (2022), archiving is not just the end of a project but ties into the broader history of research, thereby relating to other research projects. Language documentation does not take place in 'uncharted' land, since most cases link to a previous interaction with the community by scholars (or, in many cases, missionaries) and the artefacts they left behind as their 'legacy'. Historically and socially aware research needs to take these into consideration by carefully examining the legacy and contextualising new research amidst historical. A perspective that ignores previous work must be considered unethical, as it removes communities and generations of speakers (Bird, 2020). At the same time, a dialogue with the present-day communities can enhance the curation of archival deposits, e.g. through crowdsourcing or annotations by citizen scientists (Chen & Tsay, 2017). This grants agency and control over the data to the communities, while also inspiring discourse between us and the historical communities. This conscious curation reconciles

our knowledge with theirs through the stories and narratives of research we reconstruct. This perspective of reconciliation is common in case studies of linguistic legacy materials (e.g. Dobrin, 2021; Beer, 2021; Wayt, 2021; Schwartz, 2021; O'Neill & Schwartz, 2021), and visible in my own research: Not only did the curation of the Kraasna data find overlaps between different research projects, but it also connected people in a social process (e.g. the work on the Kraasna materials was received positively by the related Seto community; conversations I had with a native speaker about the recordings). The role of the archive is changing to accommodate these discourses. Accessibility is crucial to this end, whereby artefacts and archives need to be brought to the people (Dale, 2022; Khait et al., 2022) and not expect communities to come searching for artefacts. While this avenue is still relevant, knowledge and memory institutions can approach communities in academia or beyond and invite curators to work with their artefacts, e.g. to create collections or design exhibitions around them (Woodbury, 2014). In times when access and citation metrics can skew the image of research (Ma, 2021, 2022), it is important to keep a critical eye on which data sets are used and which are not – investigating how reuse of underused data can be improved through curation needs to be a focus of scholarly activity that is mindful of its past (Weber, 2021c).

3.4 Using the framework

The previous sections have elaborated on the role of artefacts, knowledge generation, and the archive as a location for both artefacts and meaning-making in the framework presented in Table 3.1. Despite the emphasis on interpretation and movements between human and non-human bearers of knowledge, the 'social dimensions' of knowledge generation, some practical questions follow from the framework. Certain methods or approaches to working with legacy materials and the design of participatory archives have been mentioned above, yet this section will consider the practical application of the framework in more detail. Overall, it is understood as a way of fostering awareness on how different domains relate to each other, internally and externally, as well as on an abstract or material plane. Discussing linguistic legacy materials in the framework involves an understanding of the history of science and hermeneutic processes in the analysis of textual artefacts that stem from the scientific endeavour. A researcher needs to be aware of historic principles of documentation, personal and collective (e.g. academic) factors influencing the concrete fieldwork situations, as well as the linguistic, sociolinguistic, and documentary perspectives on the language and its speakers. Through a careful reconstruction and reconciliation, the narrative of the fieldwork encounter that is contained in each artefact becomes visible, whereby inferences from the material plane are combined with real-world knowledge about the history of language documentation. This moves legacy materials and the archives which host them back into the focus of present-day research and knowledge generation. Research can learn from the experiences of the past and seek a connection with (or distance from) previous generations of researchers and speakers; the linguistic legacy materials are not just preserving information but set the stage for discovery and negotiation of meaning.

In practice, the methods and tools for the curation and reconstruction of linguistic legacy materials stem from different disciplines like historical science, information science, anthropology, and, importantly, philology with its strands of palaeography, textual scholarship, or editing. This is not meant as an exhaustive list of skills that form the framework and could be simply applied to any set of linguistic legacy materials. A curator needs to establish the need for either of these methods or further sources (e.g. ethnomusicology, arts) in the process of evaluating and recontextualising the materials (compare Figure 2.1). A holistic image of the documentation project and the history of thought behind each artefact arises from the combination of different domains of knowledge, thinking along the dimensions of the framework and transferring knowledge between planes. For this dissertation, the papers are all part of a domain, as can be seen in Table 3.1, yet the reconstruction within each domain required the addition of knowledge from other parts of the framework and careful yet critical analysis, whereby I am becoming part of the reconstructed versions. The application of the framework implies a discursive practice, whereby the curator enters into communication with stakeholders either through their participation and involvement in the process or through inference from artefacts (i.e. mostly texts and paratexts), in what resembles Foucault's 'archaeology of knowledge' (2010 [1969]) as a process of knowledge generation. At the core of this endeavour lie the critical edition and curation, of the artefacts and the contexts of their creation and reception, i.e. recontextualisation and reconstruction.

While the discussion above is critical of ideals of objectivity as a solution to reproducibility issues, the framework does not discredit the use of methods like randomised trials, cross-validation or double-blinding. If they are applied in a way that makes it possible to track and attribute interpretive movements between planes, this methodology is not obscuring the relevant contexts and insights into the origin of knowledge. Although the ideal of neutrality, as a position outside of human perception and cognition, cannot be reached by these approaches, they can create a basis for negotiation and discourse between people, and can yield interesting scenarios: An anonymous reviewer for Paper I asked me to check my terminology against that of other researchers in the field, among others also 'Tobias Weber'. Through the double-blind review, the roles as the author of this paper and my previous papers became separated, enabling a discourse between these roles, whereby I reconciled my terminology with my previously established wordings. Yet, this would not have been different in a single-blind review, nor would an open review discredit the discourse between the reviewer, the author, and other positions in the literature. Any of these roles can contribute to the process of meaning-making and add value to a paper, or any artefact in general, through interpretive movements. These comments are, thus, not reproducible by any other reviewer but are tied to this particular person and their individual knowledge and experiences. While the review process is meant to reduce subjective judgements and overt positionality, it cannot remove them altogether, as the communication represents a discursive act that co-constructs knowledge socially among members of the academic community. But in any processes of knowledge creation, there are subjective factors that we can analyse under the idea of ownership and attribution of human influence. This does not constitute a legal framework, yet legal concepts and definitions

of intellectual property and access rights are increasingly important for language data and linguistic legacy materials (Austin, 2010).

	abstract	**material**
micro-level	intellectual property rights	creator roles, e.g. CRediT or Tromsø recommendations
macro-level	cultural heritage	copyright and publishing licences

Table 3.2: Ownership and attribution

Table 3.2 highlights how different types of rights and attribution are of importance within the framework: Abstract knowledge is most commonly understood as intellectual property that can be contained in an artefact produced, e.g. recorded, transcribed, or deposited, by its creator. There are different metadata schemas which contain information on these contributor roles (cf. Andreassen et al., 2019). On a macro-level, these individual artefacts, songs, narratives, or broadly "creations" are linked to those of other people, e.g. texts that are part of a cultural heritage or oral tradition, which can be analysed in respect to variation and a textual genealogy. In terms of material ownership, archives tend to hold rights over their records, whereas secondary sources and paratexts fall under copyright and publishing licences from publishers and other institutions. Each of these instances adds a layer of their own interpretations, or value, to their versions of an artefact or text. Ideally, all roles are recorded and contained in thick metadata, yet it is the task of the curator to investigate and question these metadata critically to uncover contributions and positions of all involved instances. It might be tempting to reject ownership by a pseudonym but, as they form distinguishable entities, there is a way of discerning their contributions – at least, in most cases through the role of a researcher or institution that introduced them and could potentially resolve issues with metadata. In any case, an anonymised owner or creator causes issues for interpretation for more than domain-internal description. Table 3.3 presents understandings of anthroponyms within each domain, where the links across domains and planes can only be established through human interpretation and meaning-making. Internally, references to individuals, either as abstract references within a text, within a social setting, or a documentation project can fulfil their purpose by resolving to a single entity. Yet, if one decides to resolve anonymised metadata to relatives or family members, infer their social standing or attributes, or link speakers in one artefact across collections, this requires the detective work of a forensic linguist or philologist. The Kraasna legacy materials had multiple issues in this regard, where speaker names in the metadata could not be directly attributed to individuals (in times of changing name customs, see Weber, 2021b) and the archival metadata for Ojansuu's recordings list his colleague Väisänen as the creator (see Paper II and section 2.1.1).

The same efforts in reconstructing and interpreting metadata in relation to real-world knowledge is required in any step of working with scientific data. A reader of a publication

	abstract	material
micro-level	referents in a text or real-world entities, including variations of names	metadata including pseudonyms or indexical numbers
macro-level	social standing, telling names, genealogies	metadata of collections, principle of provenance

Table 3.3: Anthroponyms in the framework

must be able to understand where information and data come from and how knowledge was generated with them, i.e. which interpretive process were applied in the construction of knowledge. In a conference paper (Figure 3.1), I tried to illustrate the links between the movement between data and text, i.e. knowledge generation based on language data, and their embedding in multiple contexts. There are elements which lie further in the periphery that can influence the contexts of knowledge generation, as well as a spatiotemporal dimension that runs from top to bottom, from the contexts of creation to the contexts of reception. In the paper (Weber, 2020b), I argue that these pieces of information need to be cross-referenced and accessible from each point in the diagram; scientometric research which aims to understand the processes of knowledge generation and dissemination holistically needs to be able to investigate these factors, especially those related to human judgement and agency. The goal is not to establish a firm authoritative version of data (Seidel, 2016) but to keep contexts transparent, added value attributable, and informed decisions well-reasoned and documented. If we assume that each researcher is not just using, or even reusing, data but establishing their own version of them with traces of their individual perspective and perception, it is relevant that we are not just tracking versions but investigating how they were used to make meaning and communicate knowledge between individuals (i.e. speakers among each other, speakers and fieldworkers, researchers among each other).

In this respect, we are not facing a reproducibility crisis but an anonymity crisis: It is not possible to remove all contexts and make the human influence absolutely invisible without losing track of where knowledge comes from. Even in times of standardised procedures to guarantee anonymous review, the contextual factors, especially human ones, shape academic careers and processes of knowledge creation (Sekara et al., 2018). Considering the above example where a reviewer suggested that I consult with myself, I am certain that they were aware of the possibility that the authors they suggested could be the same authors as for the paper under review – this implicit knowledge can lead to an implicit positioning and influences the stance of negotiation and consensus forming. While I agree that certain biases may be prevented by keeping identities anonymous, there is a social dimension of knowledge construction even in the anonymised or pseudonymised discourse. The anonymity crisis thus lies in the fact that it is not possible to reduce 'truth' to data. This frequent assumption needs to critically be seen as a Western bias that does not consider the social factors and human agents who contributed to the data and artefacts, and ignores indigenous knowledge systems that posit the origin of these language data outside

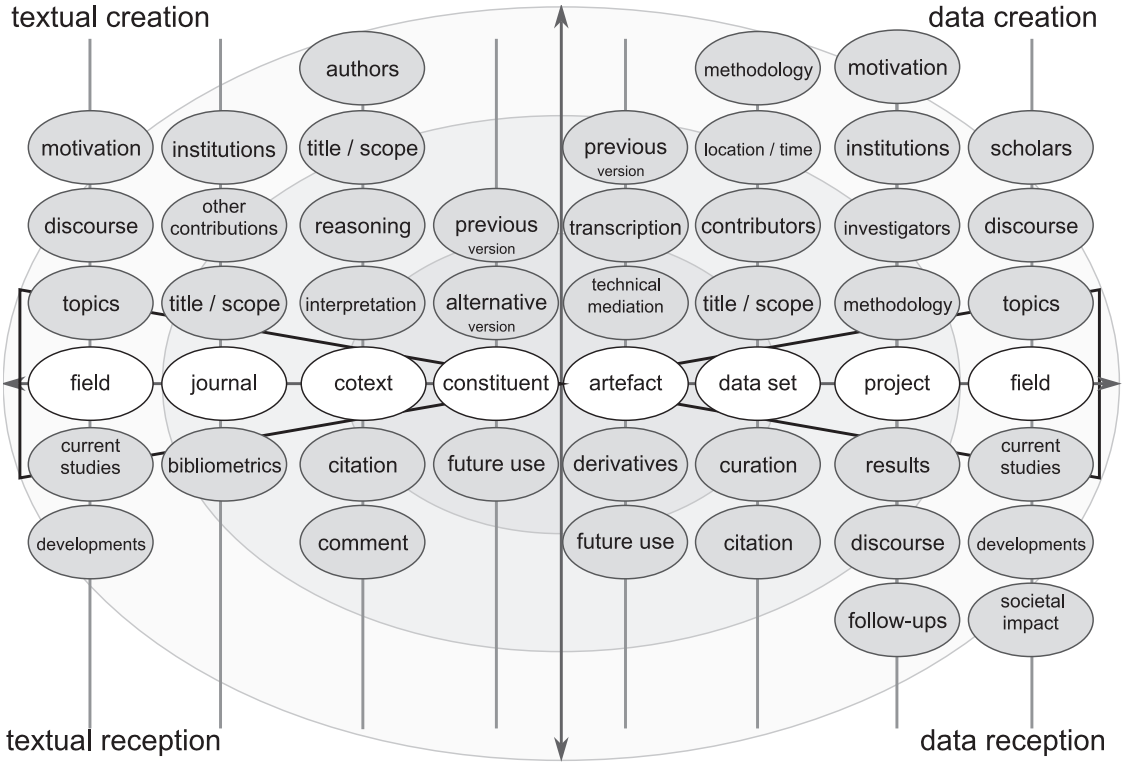

Figure 3.1: A network of data artefacts and scientific texts.
(Weber, 2020b, 230, reproduced with permission from Springer Nature)

of a single human creator. The more anonymised data and research is supposed to be, the less transparent become human factors that affect processes of knowledge creation which continue to influence the actions of individuals and knowledge and memory institutions. On the other end of the spectrum, we might face a recontextualisation crisis: If we follow the 'human in the loop', we need access to large amounts of contextual metadata and documentation of documentation, or with Geertz: 'winks upon winks upon winks' (1973, 9). This is where I consider the premise of the framework to go into a different direction than approaches emphasising the need for an increasing amount of metadata. If no amount of metadata will be sufficient to capture the situation in all details, the way of thinking about the resulting artefacts needs to change: A 'thick' description which keeps the human factors visible and allows a secondary researcher to retrace the flow of knowledge. In the interrogation of the human hand in transcription and description and the search for contextual clues about positionality or purposes behind our research, the endeavour of creating knowledge retains some transparency. We must acknowledge that there are limits to what we can perceive and record in the (meta-)documentation, making the renegotiation of meaning

a central part of the process. This must involve different communities and stakeholders, which adds perspectives and fosters trust and transparency in our analyses. If these social dimensions of our legacy as present-day researchers, i.e. linguistic legacy materials of the future, are retracable in our workflows and the contexts in which we generated knowledge are reconstructable, we enable future researchers to enter into a discourse with us through the material and abstract legacy. This work involves reflective and interpretive tasks, as a movement between the planes of the framework, through which knowledge of different generations is reconciled and insights into practices of past, present, and future generations can be gleaned.

Chapter 4

Outlook

This dissertation discusses the dynamics of knowledge generation and loss with a focus on cyclical processes surrounding linguistic legacy materials. In itself, this work constitutes an artefact in the research history on the South Estonian Kraasna variety, therefore it will be subject to the same processes of forgetting, remembering, and reconciling with new perspectives in the future. This means that the work on Kraasna is not fully done, even if the papers in this dissertation led to new and thorough descriptions of the variety and the legacy materials – the papers and this dissertation are a snapshot of knowledge at a seemingly arbitrary point in time when they were turned into artefacts. They themselves are closed systems, on the internal plane, yet fit into the larger external context of this dissertation project. The introductory chapters in this text transfer the knowledge between the papers and link them to each other, to an abstracted theoretical framework of linguistic legacy materials, and to metascientific discourses about epistemology and phenomenology. Applying the framework to this scientific work itself, it is possible to detect developments over time and reconstruct how my understanding of the Kraasna variety, linguistic legacy materials, and the hermeneutic task in their curation has changed through social embeddings. These social dimensions comprise formal and informal correspondence with colleagues and peers, reviewer comments, and impressions from conferences – an incomplete list is contained in the acknowledgements. The framework allows description and discussion of any artefact within the process of knowledge generation. This dissertation and its parts are internal and external contexts to each other, while also tied to further discourses; the transfer between the papers and external links forms the social dimension of my exegesis, as well as the readers' reconciliation with their experiences and understandings. While knowledge generation and academic discourse in respect to linguistic legacy materials, the Kraasna variety, and its artefacts will continue, there are some lessons learned through the reflexive perspective on working with the Kraasna legacy materials.

First, it is important to emphasise that the curatorial work with linguistic legacy materials and the reconstruction of a meta-documentation have as much value as documentation and description (see Paper I and Paper III). Not only are skills and knowledge necessary in this process, the added value in the curation ties the secondary researcher to their versions with all the responsibilities and merit this entails. While we should aim to move

beyond metrics and quantitative estimations of value, the curation and reconstruction of legacy materials is a slow and laborious task, albeit inspiring and thought-provoking – the transcription and comparison of the recordings with the manuscripts took over 60 hours of listening to distorted speech and the physical wear of decades on the materials. This magnitude of time for curating a data set is not unusual, as Perry & Netscher (2022) mention similar amounts of time and effort needed for preparing data sets for dissemination. In their work, they link the associated costs of data sharing to data cleaning and documentation, where the complexity of data and the amount of necessary (meta-)documentation drive costs. As linguistic legacy materials often require the reconstruction of the meta-documentation and a full recontextualisation of complex data sets, this process is not straightforward but requires human judgement and careful evaluation.

Second, in the process of curating linguistic legacy materials, I advocate for a philological stance (Paper I) that should be emphasised in academic curricula (Paper III) and supported by knowledge and memory institutions (Paper IV). This stance does not only involve the knowledge of tools and methods but refers to a reflexive and aware perspective on the processes related to knowledge creation within an academic setting and beyond. Especially when working with indigenous language data and traditional knowledge systems, it is important to include communities and stakeholders in the process. Through their participation, the curation does not only reconcile different eras of research but also generations of speakers, ideally in meaningful ways that create a value to all involved parties. In this co-construction of meaning, a holistic view is required of the necessary skills to navigate the dynamic framework and all domains within it. While expert knowledge of community members, linguists, or archivists cannot be replaced, a strong focus on the linguistic roots in the humanities is advisable over a technical perspective that reduces individuals to metadata.

Third, it is important to keep present-day linguistics as reconstructable as possible (Dobrin & Schwartz, 2021), whereby contexts of knowledge creation and reception can be restored and discourse on *how* we want to conduct research takes place. This involves epistemological as well as ontological questions, which this dissertation explored from a hermeneutic and phenomenological perspective based on the material and abstract forms of knowledge. In managing these transfers and interpretations, a range of literacies are required – while data, media, and information literacy form a very close unit of abilities (Calzada Prado & Marzal, 2013), a transfer into other domains may also involve artistic, social, or communicative skills to address the right audiences. Only if all relevant parties are involved in the social construction and negotiation of meaning, they can reach a consensus that respects and embraces different perspectives. The negotiation and reconciliation with previous generations and the decisions that led to the form of linguistic artefacts we find today can be seen as a form of *subjectivity management*, which ties in with theories of knowledge creation and scientific processes (cf. Bloor, 1991; Longino, 2002). The goal for linguistics is not to establish an authoritative version of reality or to aim for an asocial picture of language that may be reproducible but has no links to any past or present speakers. Instead, the goal is a transparent discipline that is aware of its various frameworks and traditions and tries to communicate rather than to conceal them.

Fourth, we need to question what purposes linguistic legacy materials fulfil and seek new ways of utilising and engaging with knowledge of the past. While this dissertation has focused primarily on metascientific aspects and insights into field encounters in the past, a linguist may ask what purpose an artefact serves if its contextual clues have been lost and it cannot be deciphered or interpreted. The basic phenomenological description within each domain will still be possible in this scenario, i.e. material description, comparisons, impressions. In meaning-making, the researcher draws from their experience and constructs knowledge around the artefact, yet the version of language or the description will always be tied to the assumptions they make in the progress. Linguistics can learn from the framework that multiple perspectives are commonplace not just with historical artefacts but also with those of present-day research. The difference lies in our access to contextual information and knowledge about conventions, goals, and settings of the documentation project, as well as individual human factors. There is a value to all artefacts, even drafts or test recordings – the value of an artefact might not be apparent for linguistics and its abstract analysis of language, but in the knowledge it contains about contexts and practices. If we consider that some artefacts like doodles, sketch notes, or memoranda were not meant to be preserved or analysed by anyone but the creator – like Kallas crossed out bullet points in his notes leading to several illegible pages of discarded notes – our interpretation of their meaning will always bear strong traits of our inferred knowledge and the import we ascribe to the artefact. The process of knowledge generation and legacification is dynamic. Under this premise, any artefact can lead to new knowledge, even if this happens by inference and the posterior construction of meaning, which makes the investigation of linguistic legacy materials a fruitful endeavour. Whether or not an academic discipline can gain insights relevant to its analytical frameworks does not delimit the value of the engagement and renegotiation of meaning. Yet, it requires transparency about the premises under which we operate, for example when interpreting information that was deliberately discarded into oblivion. In this respect, we must communicate the purpose of our work and requirements or limitations of curation clearly.

Ultimately, while every paper has its own conclusion and outlook with issues and questions for future research, I would like to reiterate a quote from Paper III as a central lesson: 'curation should treat historical language data with the same respect that we would want from future generations of researchers for our own language data from present-day documentation projects'. Given that our own work will be investigated and explored by future generations of researchers looking for answers about the decisions we made in knowledge creation, it is appropriate that we lead by example and show humility when assessing the legacy of previous generations, a lesson I learned in working with the Kraasna materials. We must not dismiss them for their alleged errors but aim to understand where the discord between our present-day understanding and the decisions in the past stem from – removing data or artefacts from our research for their errors might occasionally be necessary, yet always removes the efforts of speakers and researchers who were involved in the documentation and, thereby, a communicative act that tries to transcend time. In this respect, linguistic legacy data have a firm place within linguistics and we should pay attention to what we can still learn from them.

Chapter 5

Published works

Four original research papers are part of this dissertation – they can be obtained from the publishers: Papers I–III are published under Open Access agreements and freely available from the University of Virginia's Aperio service (Paper I), the University of Tartu Press (Paper II), and Ubiquity Press (Paper III); the published version of Paper IV is copyrighted by Emerald Publishing and cannot be reproduced here. All four papers are single authored and contain my original research, as part of this dissertation project.

5.1 Paper I

Weber, Tobias. 2021d. Philology in the folklore archive: Interpreting past documentation of the Kraasna dialect of Estonian. In Lise M. Dobrin & Saul Schwartz (eds.), *Language Documentation and Description* 21. Special Issue on the Social Lives of Linguistic Legacy Materials. London: EL Publishing. 70-100.
https://doi.org/10.25894/ldd18

5.2 Paper II

Weber, Tobias. 2021a. A linguistic analysis of Heikki Ojansuu's phonograph recordings of Kraasna. In Karl Pajusalu & Uldis Balodis (eds.). *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 12(2). 343-390.
https://doi.org/10.12697/jeful.2021.12.2.13

5.3 Paper III

Weber, Tobias. 2021e. The Curation of Language Data as a Distinct Academic Activity: A Call to Action for Researchers, Educators, Funders, and Policymakers. *Journal of Open Humanities Data* 7(28).
http://doi.org/10.5334/johd.51

5.4 Paper IV

Weber, Tobias. 2022. Conceptualising language archives through legacy materials. *The Electronic Library* 40(5). 525-538.
https://doi.org/10.1108/EL-02-2022-0029

Bibliography

Abraham, Terry. 1991. Oliver W. Holmes Revisited: Levels of Arrangement and Description in Practice. *The American Archivist*, 54(3): 370–377.

Altenmüller, Marlene Sophie, Leonie Lucia Lange, & Mario Gollwitzer. 2021. When research is me-search: How researchers' motivation to pursue a topic affects laypeople's trust in science. *PLOS ONE*, 16(7): 1–19.

Andreassen, Helene N., Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, & the Research Data Alliance Linguistic Data Interest Group. 2019. *Tromsø recommendations for citation of research data in linguistics*. Research Data Alliance.

Appadurai, Arjun. 1986. Introduction: commodities and the politics of value. In Arjun Appadurai (ed.). *The Social Life of Things: Commodities in Cultural Perspective*, 3–63. Cambridge University Press.

Aristar-Dry, Helen. 2009. Preserving digital language materials: Some considerations for community initiatives. In Wayne Harbert, Sally McConnell-Ginet, Amanda Miller, & John Whitman (eds.). *Language and Poverty*, 202–222. Multilingual Matters, Bristol.

Assmann, Aleida. 2008. Canon and archive. In Astrid Erll & Ansgar Nünning (eds.). *Cultural memory studies: An international and interdisciplinary handbook*, 97–107. Mouton de Gruyter, Berlin.

Assmann, Jan. 2011. *Cultural Memory and Early Civilization: Writing, Remembrance, and Political Imagination*. Cambridge University Press, Cambridge.

Austin, Peter K. 2010. Communities, ethics and rights in language documentation. *Language Documentation and Description*, 7: 34–54.

Austin, Peter K. 2013. Language documentation and meta-documentation. In Mari Jones & Sarah Ogilvie (eds.). *Keeping Languages Alive. Documentation, Pedagogy, and Revitalisation*, 3–15. Cambridge University Press, Cambridge.

Austin, Peter K. 2017. Language Documentation and Legacy Text Materials. *Asian and African Languages and Linguistics*, 11: 23–44.

Babinski, Sarah, Jeremiah Jewell, Juhyae Kim, Kassandra Haakman, Amelia Lake, Irene Yi, & Claire Bowern. 2022. How usable are digital collections for endangered languages? A review. *Proceedings of the Linguistic Society of America*, 7(1): 5219.

Balodis, Uldis & Karl Pajusalu. 2021. Introductory survey of the South Estonian language islands. *Journal of Estonian and Finno-Ugric Linguistics*, 12(2): 7–31.

Banta, Natalie M. 2016. Death and privacy in the digital age. *North Carolina Law Review*, 94: 928–990.

Beer, Samuel J. 2021. Interdisciplinary aspirations and disciplinary archives: Losing and finding John M. Weatherby's Soo data. *Language Documentation and Description*, 21: 101–139.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice, & Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1): 1–18.

Bharti, Sneha & Ranjeet Kumar Singh. 2022. Evaluation and analysis of digital language archives development platforms: a parametric approach. *The Electronic Library*, 40(5): 552–567.

Bird, Steven. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, 3504–3519, Barcelona. International Committee on Computational Linguistics.

Blokland, Rogier, Niko Partanen, & Michael Rießler. 2021. This is thy brother's voice : documentary and metadocumentary linguistic work with a folklore recording from the Nenets-Komi contact area. In Mika Hämäläinen, Niko Partanen, & Khalid Alnajjar (eds.). *Multilingual Facilitation*, 208–227. University of Helsinki Library, Helsinki.

Bloor, David. 1991. *Knowledge and Social Imagery*. University of Chicago Press, Chicago.

Bond, George C. 1990. Fieldnotes: Research in Past Occurrences. In Roger Sanjek (ed.). *Fieldnotes. The Makings of Anthropology*, 273–289. Cornell University Press, Ithaca and London.

Brázdil, Rudolf, Christian Pfister, Heinz Wanner, Hans von Storch, & Jürg Luterbacher. 2005. Historical climatology in Europe – the state of the art. *Climatic Change*, 70: 363–430.

Buitelaar, J.C. 2017. Post-mortem privacy and informational self-determination. *Ethics and Information Technology*, 19: 129–142.

Calzada Prado, Javier & Miguel Ángel Marzal. 2013. Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri*, 63(2): 123–134.

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, & Maui Hudson. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1): 43.

Chen, Chih-Ming & Ming-Yueh Tsay. 2017. Applications of collaborative annotation system in digital curation, crowdsourcing, and digital humanities. *The Electronic Library*, 35(6): 1122–1140.

Clifford, James. 1990. Notes on (field)notes. In Roger Sanjek (ed.). *Fieldnotes. The Makings of Anthropology*, 47–70. Cornell University Press, Ithaca and London.

Dale, Merrion. 2022. Creating workflow for mediated archiving in CoRSAL. *The Electronic Library*, 40(5): 568–578.

Derrida, Jacques. 1973 [1967]. *Speech and Phenomena*. Northwestern University Press, Evanston.

Derrida, Jacques. 1995. A Freudian impression. *Diacritics*, 25(2): 9–63.

Derrida, Jacques. 1997. *Of Grammatology*. Johns Hopkins University Press, Baltimore and London.

Dilthey, Wilhelm. 1990 [1883]. *Einleitung in die Geisteswissenschaften. Versuch einer Grundlegung für das Studium der Gesellschaft und der Geschichte*, volume 1 of *Wilhelm Dilthey. Gesammelte Schriften*. Vandenhoeck & Ruprecht, Göttingen.

Dobrin, Lise M. 2012. Ethnopoetic analysis as a methodological resource for endangered language linguistics: The social production of an Arapesh text. *Anthropological Linguistics*, 54(1): 1–32.

Dobrin, Lise M. 2021. The Arapesh "suitcase miracle": The interpretive value of reproducible research. *Language Documentation and Description*, 21: 37–69.

Dobrin, Lise M. & Saul Schwartz. 2021. The social lives of linguistic legacy materials. *Language Documentation and Description*, 21: 1–36.

Douglas, Jennifer. 2010. Origins: evolving ideas about the principle of provenance. In Terry Eastwood & Heather MacNeil (eds.). *Currents of Archival Thinking*, 23–43. Libraries Unlimited, Santa Barbara.

Douglas, Ty-Ron. 2017. My Reasonable Response: Activating Research, MeSearch, and WeSearch to Build Systems of Healing . *Critical Education*, 8(2): 21–30.

Duarte, Marisa Elena & Miranda Belarde-Lewis. 2015. Imagining: Creating spaces for indigenous ontologies. *Cataloging & Classification Quarterly*, 53(5–6): 677–702.

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In *Essentials of Language Documentation*, 31–66. De Gruyter Mouton, Berlin and New York.

Ehlich, Konrad. 2007. Textualität und Schriftlichkeit. In Ludwig Morenz & Stefan Schorch (eds.). *Was ist ein Text? Alttestamentliche, ägyptologische und altorientalistische Perspektiven*, 3–17. De Gruyter, Berlin and Boston.

Engh, Jan. 2006. *Norwegian examples in international linguistics literature. An inventory of defective documentation*. Universitetsbiblioteket i Oslo, Oslo.

Ernits, Enn. 2012. Fr. R. Kreutzwald lõunaeestlaste piire kompimas. In Sullõv Jüvä (ed.). *Õdagumeresoomõ piiriq*, volume 26 of *Võro Instituudi Toimõndusõq*, 30–65. Võro Instituut, Võru.

Ernits, Enn. 2018. Kraasna rahvalaulude esimestest üleskirjutustest. In Sullõv Jüvä (ed.). *Valitsõmisjaotusõst keeleaoluuni*, volume 33 of *Võro Instituudi Toimõndusõq*, 157–201. Võro Instituut, Võru.

Ernits, Enn. 2021. Kraasna nominal derivation. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 12(2): 313–341.

Everett, Daniel L. 2001. Monolingual field research. In Paul Newman & Martha Ratliff (eds.). *Linguistic Fieldwork*, 166–188. Cambridge University Press.

Foucault, Michel. 2010 [1969]. *Archaeology Of Knowledge*. Vintage Books, New York.

Fynsk, Christopher. 2003. Lascaux and the question of origins. *POIESIS: A Journal of the Arts and Communication*, 5: 6–19.

Gadamer, Hans-Georg. 1975 [1960]. *Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik*. J.C.B. Mohr, Tübingen.

Gardner, Susan K., Jeni Hart, Jennifer Ng, Rebecca Ropers-Huilman, Kelly Ward, & Lisa Wolf-Wendel. 2017. "me-search": Challenges and opportunities regarding subjectivity in knowledge construction. *Studies in Graduate and Postdoctoral Education*, 8(2): 88–108.

Geertz, Clifford. 1973. Thick Description: Toward an Interpretive Theory of Culture. In *The interpretation of cultures*, 3–30. Basic Books, New York.

Genette, Gérard. 1997. *Paratexts. Thresholds of Interpretation*. Cambridge University Press, Cambridge.

Good, Jeff. 2007. The ecology of documentary and descriptive linguistics. *Language Documentation and Description*, 4: 38–57.

Gorichanaz, Tim. 2017. There's no shortcut: Building understanding from information in ultrarunning. *Journal of Information Science*, 43(5): 713–722.

Gorichanaz, Tim. 2019. A first-person theory of documentation. *Journal of Documentation*, 75(1): 190–212.

Gourlay, Stephen. 2006. Conceptualizing knowledge creation: A critique of Nonaka's theory. *Journal of Management Studies*, 43(7): 1415–1436.

Greenberg, Jane. 2009. Theoretical considerations of lifecycle modeling: An analysis of the Dryad Repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging & Classification Quarterly*, 47(3–4): 380–402.

Gurd, Sean. 2015. *Philology and Greek Literature*. Oxford Handbook Topics in Classical Studies.

Haendel, Melissa A., Nicole A. Vasilevsky, & Jacqueline A. Wirz. 2012. Dealing with data: A case study on information and data management literacy. *PLOS Biology*, 10(5): e1001339.

Halbwachs, Maurice. 1980. *The Collective Memory*. Harper and Row, New York.

Harris, Roy. 1980. *The language makers*. Cornell University Press, Ithaca.

Heidegger, Martin. 2018 [1927]. *Sein und Zeit*. Martin Heidegger Gesamtausgabe. Vittorio Klostermann, Frankfurt.

Henke, Ryan & Andrea L. Berez-Kroeker. 2016. A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation & Conservation*, 10: 411–457.

Higgins, Sarah. 2018. Digital curation: the development of a discipline within information science. *Journal of Documentation*, 74(6): 1318–1338.

Himmelmann, Nikolaus P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation & Conservation*, 6: 187–207.

Hjørland, Birger. 2005. Empiricism, rationalism and positivism in library and information science. *Journal of Documentation*, 61(1): 130–155.

Holton, Gary. 2014. Mediating language documentation. *Language Documentation and Description*, 12: 37–52.

Husserl, Edmund. 2002 [1913]. *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie: Allgemeine Einführung in die reine Phänomenologie*. De Gruyter, Berlin and Boston.

Huvila, Isto. 2008. Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management. *Archival Science*, 8: 15–36.

Huvila, Isto. 2015. The unbearable lightness of participating? revisiting the discourses of "participation" in archival literature. *Journal of Documentation*, 71(2): 358–386.

Iva, Sulev. 2015. Liivi, Ludzi ja Kraasna maarahva kiil. *Keel ja Kirjandus*, 7: 515–517.

Jackson, Jean E. 1990. "I am a fieldnote": Fieldnotes as a symbol of professional identity. In Roger Sanjek (ed.). *Fieldnotes. The Makings of Anthropology*, 3–33. Cornell University Press, Ithaca and London.

Kallas, Oskar. 1902. Übersicht über das sammeln estnischer runen. *Finnisch-ugrische Forschungen*, 2(1): 8–41.

Kallas, Oskar. 1903. *Kraasna maarahvas*. Soome Kirjanduse Selts, Helsinki.

Kenfield, Ayla Stein, Liz Woolcott, Santi Thompson, Elizabeth Joan Kelly, Ali Shiri, Caroline Muglia, Kinza Masood, Joyce Chapman, Derrick Jefferson, & Myrna E. Morales. 2022. Toward a definition of digital object reuse. *Digital Library Perspectives*, 38(3): 378–394.

Khait, Ilya, Leonore Luksch, & Mandana Seyfeddinipur. 2022. Linguistic archives and language communities questionnaire: establishing (re-)use criteria. *The Electronic Library*, 40(5): 539–551.

Koltay, Tibor. 2015. Data literacy: in search of a name and identity. *Journal of Documentation*, 71(2): 401–415.

Kroskrity, Paul V. 2015. Discursive discriminations in the representation of Western Mono and Yokuts stories: Confronting narrative inequality and listening to Indigenous voices in Central California. In Paul V. Kroskrity & Anthony K. Webster (eds.). *The legacy of Dell Hymes: Ethnopoetics, narrative inequality, and voice*, 135–163. Indiana University Press, Bloomington.

Kuhn, Thomas S. 2012 [1962]. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Kukk, Armin. 1938. *Kraasna murde esimese silbi vokalism*. Term paper at the University of Tartu. Archived in the Archives of Estonian Dialects and Kindred Languages (H0168).

Kurtz, Mary. 2010. Dublin Core, DSpace, and a brief analysis of three university repositories. *Information Technology and Libraries*, 29(1): 40–46.

Latour, Bruno. 1987. *Science in action: how to follow scientists and engineers through society*. Harvard University Press, Cambridge.

Latour, Bruno. 1993 [1991]. *We have never been modern*. Harvard University Press, Cambridge.

Lehmann, Christian. 2004. Data in linguistics. *The Linguistic Review*, 21(3-4): 175–210.

Leroi-Gourhan, André. 1993. *Gesture and Speech*. MIT Press, Cambridge and London.

Linguistic Society of America. 1994. *The Need for the Documentation of Linguistic Diversity*. Statement, 1 June 1994.

Linguistic Society of America. 2010. *Recognizing the Scholarly Merit of Language Documentation*. Resolution, 8 January 2010.

Linguistic Society of America. 2018. *Statement on Evaluation of Language Documentation for Hiring, Tenure, and Promotion*. 25 September 2018.

Lohmar, Dieter. 1997. Truth. In Lester Embree, Elizabeth A. Behnke, David Carr, J. Claude Evans, José Huertas-Jourda, Joseph J. Kockelmans, William R. McKenna, Algis Mickunas, Jitendra Nath Mohanty, Thomas M. Seebohm, & Richard M. Zaner (eds.). *Encyclopedia of Phenomenology*, 708–712. Springer Netherlands, Dordrecht.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton.

Longino, Helen E. 2002. *The Fate of Knowledge*. Princeton University Press, Princeton.

Lukin, Karina. 2017. Matthias Alexander Castrén's notes on Nenets folklore. *Suomalais-Ugrilaisen Seuran Aikakauskirja*, 96: 169–211.

Ma, Lai. 2021. Metrics as time-saving devices. In Filip Vostal (ed.). *Inquiring into Academic Timescapes*, 123–133. Emerald Publishing, Bingley.

Ma, Lai. 2022. Metrics and epistemic injustice. *Journal of Documentation*, 78(7): 392–404.

Marcus, George E. & Dick Cushman. 1982. Ethnographies as texts. *Annual Review of Anthropology*, 11: 25–69.

Markus, Lynne M. 2001. Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems*, 18(1): 57–93.

Marshall, Catherine C. 1997. Annotation: From paper books to the digital library. In *Proceedings of the Second ACM International Conference on Digital Libraries*, DL '97, 131–140, New York. Association for Computing Machinery.

McGann, Jerome. 2013. Philology in a New Key. *Critical Inquiry*, 39(2): 327–346.

Meister Ko. Freitag, Raquel. 2022. Sociolinguistic repositories as asset: challenges and difficulties in Brazil. *The Electronic Library*, 40(5): 607–622.

Mets, Mari, Anu Haak, Triin Iva, Grethe Juhkason, Mervi Kalmus, Miina Norvik, Karl Pajusalu, Pire Teras, Tuuli Tuisk, & Lembit Vaba. 2014. *Lõunaeesti keelesaarte tekstid. Eesti murded*, volume IX. Eesti Keele Instituut & Tartu Ülikool, Tallinn.

Moore, Robert. 2013. Reinventing ethnopoetics. *Journal of Folklore Research*, 50(1-3).

Moore, Robert E. 2009. From performance to print, and back: Ethnopoetics as social practice in Alice Florendo's corrections to "Raccoon and his Grandmother". *Text & Talk*, 29(3): 295–324.

Muñoz, Trevor. 2013. Data Curation as Publishing for the Digital Humanities. *Journal of Digital Humanities*, 2(3).

Nancy, Jean-Luc. 1993. *The Birth to Presence*. Stanford University Press, Stanford.

Nancy, Jean-Luc. 1997. *The Sense of the World*. University of Minnesota Press, Minneapolis.

Nathan, David. 2010. Archives and language documentation: From disk space to MySpace. *Language Documentation and Description*, 7: 172–208.

Nathan, David. 2014. Access and accessibility at ELAR, an archive for endangered languages documentation. *Language Documentation and Description*, 12: 187–208.

Nathan, David & Peter K. Austin. 2004. Reconceiving metadata: language documentation through thick and thin. *Language Documentation and Description*, 2: 179–188.

Nevins, Eleanor M. 2013. *Lessons from Fort Apache: Beyond language Endangerment and Maintenance*. Wiley-Blackwell, Malden.

Nevins, Eleanor M. 2015. "grow with that, walk with that": Hymes, dialogicality, and text collections. In Paul V. Kroskrity & Anthony K. Webster (eds.). *The legacy of Dell Hymes: Ethnopoetics, narrative inequality, and voice*, 71–107. Indiana University Press, Bloomington.

Nichols, Johanna. 1998. The origin and dispersal of languages: Linguistic evidence. In Nina Jablonski & Leslie C. Aiello (eds.). *The Origin and Diversification of Language*, 127–170. San Francisco: California Academy of Sciences, San Francisco.

Nigol, Lembit. 1940. *Ühesilbiste pikavokaalsete tüvede ainsuse illatiivist Lõuna-Eesti kagumurrakutes ja keelesaartel. Ülemastme seminaritöö. EKA 590*. Term paper at the University of Tartu. Archived in the Archives of Estonian Dialects and Kindred Languages (M0151).

Niu, Jinfang. 2013. Recordkeeping metadata and archival description: a revisit. *Archives & Manuscripts*, 41(3): 203–215.

Norvik, Miina, Uldis Balodis, Valts Ernštreits, Gunta Kļava, Helle Metslang, Karl Pajusalu, & Eva Saar. 2021. The South Estonian language islands in the context of the Central Baltic area. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 12(2): 33–72.

Ojansuu, Heikki. 1912. Ein südestnischer beitrag zur stufenwechseltheorie. *Finnisch-ugrische Forschungen*, 12: 147–149.

O'Neill, Sean & Saul Schwartz. 2021. Recirculating and revitalizing words: Lexical legacies in Native American language preservation. *Language Documentation and Description*, 21: 199–228.

Pajusalu, Karl. 2007. Estonian dialects. In Mati Erelt (ed.). *Estonian language*, Linguistica Uralica Supplementary Series, 231–272. Eesti Teaduste Akadeemia, Tallinn.

Pajusalu, Karl. 2022. Seto South Estonian. In Marianne Bakró-Nagy, Johanna Laakso, & Elena Skribnik (eds.). *The Oxford Guide to the Uralic Languages*, 367–379. Oxford University Press, Oxford.

Pajusalu, Karl, Tiit Hennoste, Ellen Niit, Peeter Päll, & Jüri Viikberg. 2018. *Eesti murded ja kohanimed*. Eesti Keele Sihtasutus, Tallinn.

Palgi, R. 1937. *Häälikulooline ülevaade klusiilide ja nende nõrkade vastete esindusest Kraasna murdes*. Term paper at the University of Tartu. Archived in the Archives of Estonian Dialects and Kindred Languages (H0165).

Pasquetto, Irene V., Christine L. Borgman, & Morgan F. Wofford. 2019. Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review*, 1(2).

Pasquetto, Irene V., Bernadette M. Randles, & Christine L. Borgman. 2017. On the reuse of scientific data. *Data Science Journal*, 16: 8.

Perreault, Charles & Sarah Mathew. 2012. Dating the origin of language using phonemic diversity. *PLoS One*, 7(4): e35289.

Perry, Anja & Sebastian Netscher. 2022. Measuring the time spent on data curation. *Journal of Documentation*, 78(7): 282–304.

Petőfi, János S. 1973. Text-grammars, text-theory and the theory of literature. *Poetics*, 2(3): 36 – 76.

Poulos, Christopher N. 2021. *Essentials of autoethnography*. American Psychological Association.

Ross, Seamus. 2012. Digital preservation, archival science and methodological foundations for digital libraries. *New Review of Information Networking*, 17(1): 43–68.

Sanjek, Roger. 1990. A vocabulary for fieldnotes. In Roger Sanjek (ed.). *Fieldnotes. The Makings of Anthropology*, 92–121. Cornell University Press, Ithaca and London.

Schwartz, Saul. 2021. Legacy materials and cultural facework: Obscenity and bad words in Siouan language documentation. *Language Documentation and Description*, 21: 166–198.

Seidel, Frank. 2016. Documentary linguistics: A language philology of the 21st century. *Language Documentation and Description*, 13: 23–63.

Sekara, Vedran, Pierre Deville, Sebastian E. Ahnert, Albert-László Barabási, Roberta Sinatra, & Sune Lehmann. 2018. The chaperone effect in scientific publishing. *Proceedings of the National Academy of Sciences*, 115(50): 12603–12607.

Setälä, Emil Nestor. 1901. Über die transskription der finnisch-ugrischen sprachen. *Finnisch-ugrische Forschungen*, 1: 15–52.

Thessen, Anne E., Matt Woodburn, Dimitrios Koureas, Deborah Paul, Michael Conlon, David P. Shorthouse, & Sarah Ramdeen. 2019. Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations of the RDA/TDWG Working Group. *Data Science Journal*, 18: 54.

Thibodeau, Kenneth. 2001. Building the archives of the future. advances in preserving electronic records at the National Archives and Records Administration. *D-Lib Magazine*, 7(2).

Thieberger, Nicholas & Michel Jacobson. 2010. Sharing data in small and endangered languages: Cataloging and metadata, formats, and encodings. In Lenore A. Grenoble & N. Louanna Furbee (eds.). *Language documentation: Practice and values*, 147–158. John Benjamins, Amsterdam.

Topp, Warren. 2000. Generative conversations: applying Lyotard's discourse model to knowledge creation within contemporary organizations. *Systems Research and Behavioral Science*, 17(4): 333–340.

Tsoukas, Haridimos. 2009. A dialogical approach to the creation of new knowledge in organizations. *Organization Science*, 20(6): 941–957.

Tuhiwai Smith, Linda. 1999. *Decolonizing Methodologies: Research and Indigenous Peoples.* Zed Books, London.

Van Manen, Max. 2014. *Phenomenology of Practice: Meaning-Giving Methods in Phenomenological Research and Writing.* Routledge.

Vogel, Coleen H. 1989. A documentary-derived climatic chronology for South Africa, 1820–1900. *Climatic Change*, 14: 291–307.

Wasson, Christina. 2021. *Participatory design of language and culture archives.* Oxford Research Encyclopedia of Anthropology.

Wasson, Christina, Gary Holton, & Heather S. Roth. 2016. Bringing User-Centered Design to the Field of Language Archives. *Language Documentation & Conservation*, 10: 641–681.

Wayt, Josh. 2021. Documenting language and discerning listenership: Fluent speakers' evaluations of Dakota's oldest legacy texts. *Language Documentation and Description*, 21: 140–165.

Weber, Tobias. 2020a. Metadata Inheritance: New Research Paper, New Data, New Metadata? In *Reframing Research Workshop Accepted Papers.* Zenodo.

Weber, Tobias. 2020b. A philological perspective on meta-scientific knowledge graphs. In *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*, 226–233, Cham. Springer International Publishing.

Weber, Tobias. 2021a. A linguistic analysis of Heikki Ojansuu's phonograph recordings of Kraasna. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 12(2): 343–390. [**Paper II of this thesis**].

Weber, Tobias. 2021b. Consultant Identity in Historical Language Data: Anthroponyms as a Tool or as an Obstacle? In *Proceedings of the International Onomastic Conference "Anthroponyms and Anthroponymic Researches in the Beginning of 21st Century"*, 165–175, Sofia. Bulgarian Academy of Sciences.

Weber, Tobias. 2021c. Mind the gap: Language data, their producers, and the scientific process. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *OpenAccess Series in Informatics (OASIcs)*, 6:1–6:9, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Weber, Tobias. 2021d. Philology in the folklore archive: Interpreting past documentation of the Kraasna dialect of Estonian. *Language Documentation and Description*, 21: 70–100. [**Paper I of this thesis**].

Weber, Tobias. 2021e. The Curation of Language Data as a Distinct Academic Activity: A Call to Action for Researchers, Educators, Funders, and Policymakers. *Journal of Open Humanities Data*, 7(28). [**Paper III of this thesis**].

Weber, Tobias. 2022. Conceptualising language archives through legacy materials. *The Electronic Library*, 40(5): 525–538. [**Paper IV of this thesis**].

Weber, Tobias & Mia Klee. 2020. Agency in scientific discourse. *Bulletin of the Transilvania University of Braşov Series IV: Philology and Cultural Studies*, 13(1): 71–86.

Wendell Compton, Bradley. 2014. Ontology in information studies: without, within, and withal knowledge management. *Journal of Documentation*, 70(3): 425–442.

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, & Barend Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1): 160018.

Winberry, Joseph & LaVerne Gray. 2022. From "Mesearch" to "Wesearch": The role of community in developing Identity-Centric Research. In *Proceedings of the Association for Library and Information Science Education. Annual Conference: ALISE 2022. Go Back and Get It: From One Narrative to Many*, Urbana-Champaign. University of Illinois.

Windfeld Lund, Niels. 2010. Document, text and medium: concepts, theories and disciplines. *Journal of Documentation*, 66(5): 734–749.

Woodbury, Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. *Language Documentation and Description*, 12: 19–36.

Wu, Anping & Jiangping Chen. 2022. Sustaining multilinguality: case studies of two multilingual digital libraries. *The Electronic Library*, 40(6): 625–645.

Xie, Iris, Rakesh Babu, Shengang Wang, Hyun Seung Lee, & Tae Hee Lee. 2022. Assessment of digital library design guidelines to support blind and visually impaired users: a study of key stakeholders' perspectives. *The Electronic Library*, 40(6): 646–661.

Yi, Irene, Amelia Lake, Juhyae Kim, Kassandra Haakman, Jeremiah Jewell, Sarah Babinski, & Claire Bowern. 2022. Accessibility, discoverability, and functionality: An audit of and recommendations for digital language archives. *Journal of Open Humanities Data*, 8: 10.

Yoon, Ayoung. 2017. Role of communication in data reuse. *Proceedings of the Association for Information Science and Technology*, 54(1): 463–471.

Acknowledgements

I want to express my gratitude to all colleagues, peers, and institutions who supported me in my research and shaped my personal research trajectory with enriching thoughts, feedback on my work, and personal support. I am deeply indebted to Elena Skribnik as my primary supervisor, whose unwavering belief in my research guided me from my Bachelor studies to the completion of this doctoral project. I am also extremely grateful to Ksenia Shagal who joined my PhD project as a second supervisor and supplied many great ideas on the final shape of the dissertation. Furthermore, I would like to express my deepest appreciation to colleagues and previous supervisors with whom I had many long conversations and correspondences about Kraasna and linguistic legacy materials: Jeremy Bradley, Peter K. Austin, Karl Pajusalu, and Lise Dobrin. The latter also supported the publication of my papers in their collections, which laid the foundation for this project. In this respect, I am also thankful to the other guest editors and peers who guided me through the publication process: Saul Schwartz (Paper I), Uldis Balodis (Paper II), Richard Griscom, Lauren B. Collister and Hugh J. Paterson III (Paper III), and Oksana Zavalina and Shobhana L. Chelliah (Paper IV).

Thanks should also go to the 10 anonymous reviewers for sharing their wealth of experience, knowledge, and insights from different disciplines and academic traditions – my papers bear witness to their diligent service to the academic community. The same goes for all organisers, reviewers, and commentators at the conferences where I presented this research and gained important insights into the different target audiences for this framework. Special mention and deepest thanks in this respect shall be relayed to the Graduate School Language & Literature, its members and doctoral students, and the academic community in Uralic studies including the staff at the Institute for Finno-Ugric Studies at LMU and the strategic partnerships of INFUSE and COPIUS. I would be remiss in not emphasising the role of Sulev Iva (Jüvä Sullõv) as my teacher for South Estonian, who helped me with any questions about this language. On a personal note, this endeavour would not have been possible without the support of my family, as well as my close friends and fellow students who always listened to my thoughts and ideas in the most attentive and interested manner, and offered their support. Among those, I would like to acknowledge the help of Sorcha Hazelton who read numerous drafts of my manuscripts, as well as Mia Klee for hours of discussion about the Kraasna legacy materials.

Lastly, I would like to recognise the support I received from the archives hosting the Kraasna legacy materials: the Estonian Folklore Archives of the Estonian Literary Museum

in Tartu, the University of Tartu Archives of Estonian Dialects and Kindred Languages, the Archive of the Estonian Dialects and Finno-Ugric Languages at the Institute of the Estonian Language, the literary archive at the Finnish Literary Society, the sound archive of the Finnish Literary Society, as well as the Kalevala Society and the KOTUS archives. Their staff was extremely helpful in locating and accessing resources, while also allowing me some insights into archival work. Therefore, I would also like to mention Oskar Kallas, Heikki Ojansuu, Paulopriit Voolaine and his colleagues working on Kraasna, who enabled me see the Kraasna community and its language through their eyes and their documentation projects. By extension, I am also thankful to all archivists and editors of the data who initiated and continued the construction of meaning in this case. Lastly, the entire dissertation project would not exist if for the consultants' trust and belief in the endeavour of documenting and relaying an impression of the Kraasna community and its language. While they will not be able to read this dissertation themselves, I hope that my work does justice to their efforts by honouring their legacy.