Modeling the Lifecycle of Knowledge Artefacts in Qualitative Research Methodologies

Alejandro Adorjan[1], Genoveva Vargas-Solar[2] and Regina Motz[3]

[1]*University ORT Uruguay, Uruguay*
[2]*CNRS, Univ Lyon, INSA Lyon, UCBL, LIRIS, UMR5205, F-69221, France*
[3]*Instituto de Computación (INCO) Facultad de Ingeniería, Universidad de la República, Uruguay*

Abstract

This paper proposes a knowledge artefact lifecycle model tailored for qualitative research projects. Knowledge artefacts are deep analysis artefacts that give semantic to various types of content, such as surveys, interviews, codebooks, and field diaries. The proposal emerges from an empirical study of an e-social science research project that applies Grounded Theory as its theoretical framework. In this context, we propose a model to delineate the lifecycle of artefacts within a qualitative research process workflow. Our research contributes to developing a first approach on the problem of modeling the very diverse content produced through qualitative research processes. We argue that modeling the lifecycle of knowledge artefacts within a qualitative research process could shed light on the reliability and transparency of the process itself.

Keywords

research data curation, modeling knowledge artefacts, qualitative research

1. Introduction

Constructing knowledge involves processing information, experiences, values and perceptions [1]. Qualitative research to produce knowledge often generates diverse data, including interviews, field notes, codebooks, and surveys, each requiring unique storage, retrieval, and analysis methods. The management of the lifecycle of these research artefacts, encompassing creation, storage, retrieval, sharing, and reuse, poses significant challenges due to several key factors. Firstly, the volume and variety of qualitative data require specialized storage solutions to accommodate different data types and ensure seamless retrieval and analysis. Secondly, qualitative data is inherently context-sensitive and requires meticulous documentation to preserve its contextual integrity. This sensitivity complicates the storage and retrieval process, as maintaining contextual information is essential for accurate interpretation. Moreover, the qualitative research process is often iterative and evolving, demanding flexibility to adapt to new insights and changing research questions. Ensuring the reusability and traceability of research artefacts further exacerbates these challenges.

AMW 2024: 16th Alberto Mendelzon International Workshop on Foundations of Data Management, September 30th–October 4th, 2024, Mexico City, Mexico

*Corresponding author.

†These authors contributed equally.

✉ adorjan@ort.edu.uy (A. Adorjan); genoveva.vargas-solar@cnrs.fr (G. Vargas-Solar); rmotz@fing.edu.uy (R. Motz)

Research artefacts must be thoroughly documented and version-controlled to facilitate their use in future studies. This requires rigorous versioning strategies and comprehensive metadata management, which are time-consuming and technically demanding for qualitative researchers. In this context, this article explores these issues. It proposes an initial model for conceptualising qualitative research artefacts and their lifecycle, to improve their management in ongoing research and facilitate future reuse. The lifecycle of qualitative research artefacts and data curation are closely interconnected to ensure effective data management from creation to reuse, thereby enhancing the quality and integrity of research over time.

Research data curation is preserving, preparing, and sharing artefacts [2]. Current data curation models, such as CURARE [3, 4, 5], focus on facilitating the exploration of raw data collections to assist data scientists. These studies predominantly address strategies for curating quantitative data and establishing metadata models to support data curation processes. However, there is a notable gap in the literature regarding modelling the curation process of qualitative research data. Our research aims to fill this gap by explicitly addressing the curation of qualitative research artefacts. We propose a focus on modeling the lifecycle that emerges from qualitative research processes to effectively enhance the understanding and management of these artefacts.

For research to be trustworthy and transparent, its results and research artefacts are published in open-access repositories. Qualitative research data repositories like QDA (Qualitative Data Repository)[1], typically require comprehensive metadata. This includes a detailed description of the study, outlining the research objective, methods used, geographical context, and details about the research team. Metadata must also specify dataset particulars, such as collection dates, types of data (interviews, field notes), number of participants, and access conditions. Additionally, it covers data format (text, audio, video), required software for access, and licensing terms outlining usage rights and restrictions. Furthermore, metadata should reference the publication where the research process is narratively explained. However, the clarity, accuracy and traceability of the research findings presented exclusively in the final report become the editor's responsibility. To be independent of the editors' writing, we propose representing knowledge acquisition in a specific research artefact, the *Knowledge Artefact* (KA).

This proposal arises from an empirical study of an anthropology e-social science research project utilizing Grounded Theory (GT) as the primary qualitative methodology. Our main contribution is the conceptualization of the lifecycle of knowledge artefacts within the qualitative research workflow.

The remainder of the paper is structured as follows. Section 2 discusses the background. Section 3 introduces research process workflow model representing the knowledge artefact life cycle. Section 4 demonstrates the use of the model through an experimental use case. Finally, Section 5 concludes the paper and discusses future work.

2. Background

Reproducible research is a term that encompasses scientific quality in quantitative research [6]. On the other hand, rigor in qualitative research is based on transparency, credibility, reliability,

[1] https://qdr.syr.edu

and reflexivity. Qualitative research explores human aspects and new phenomena, capturing individuals' thoughts, feelings, or interpretations of meaning and process [7].

Curation is a broad topic discussed in several knowledge areas [8]. To encourage research data sharing, data curation is essential to make data reusable and interpretable [9]. Research data curation is identifying, systematizing, managing and versioning research artefacts generated across various stages of a research project [10]. Data sharing and reuse are becoming the norm in quantitative research [11]. Current approaches to Data Curation include manual, automated, and hybrid human-machine methods [12]. Curating artefacts improves reliability and transparency [13]. Data collections' curation and exploration proposals with a quantitative perspective as CURARE [3, 4, 5] and from a qualitative standpoint [14] have been reported in the literature.

Curated data collections have the potential to drive scientific progress [15] and add transparency to the research process itself, for example, publishing reports and data collections on Dataverses. Dataverses provide a structured, flexible environment to store, access, share, and analyse datasets [16]. These repositories allow researchers to deposit, reproduce and share research artefacts [17, 18].

QDA (Qualitative Data Repository) [2] is an example of these services, that offers, a platform for managing, planning, formatting and depositing research data. Data Curation Network (DCN) provide a curation platform for projects, offering a collaborative platform for data curation across a network of repositories to advance open research by making data more ethical, reusable, and understandable [19]. DCN[3] also propose a workflow of DC steps for researchers to prepare and share research data. Digital Curation Centre (DCC) [4] presents a Curation Lifecycle Model of stages required for data curation [20, 21], one of the most influential curation lifecycles [22, 23, 24, 8].

Researchers and practitioners from qualitative areas perform research data interaction of their artefacts using annotations, labelling, querying, audio and video transcription, pattern discovery, and report generation through Computer-Assisted Qualitative Data Analysis Software (CAQDAS). CAQDAS are a well known tool for qualitative researchers, for example NVivo [5], MaxQDA[6], Taguette [7] or AtlasTI[8].

3. General approach

This section outlines the proposal's main topics, context, and approach. We begin by defining a Knowledge Artefact (KA). Next, we present Grounded Theory as a qualitative research methodology. Following this, we describe five main stages of the qualitative research process. Finally, we describe the representation of a qualitative research process workflow model.

[2] https://qdr.syr.edu

[3] https://hdl.handle.net/11299/241454

[4] https://www.dcc.ac.uk/guidance/curation-lifecycle-model

[5] https://lumivero.com/products/nvivo

[6] https://www.maxqda.com

[7] https://www.taguette.org

[8] https://atlasti.com

Knowledge Artefact Knowledge Artefact (KA) is defined as a physical material that is collaboratively created, maintained and used to support knowledge-oriented social processes [25] and knowledge-related processes within a community [26]. A KA in a qualitative context refers to interpreting several elements, such as surveys, interviews, codebooks, and field diaries. It can also include descriptions of data harvesting protocols, objectives, and analysis protocols. Additionally, research artefacts may specify the theoretical framework, research questions, bibliography, corpora, or categorical analysis results.

Grounded Theory Grounded Theory (GT) is a qualitative research methodology for studying social phenomena that enables theory development based on evidence [27, 28, 29]. Several GT approaches are presented in the literature [27, 28, 29, 27, 30]. GT is increasingly being used to investigate the human and social aspects of Software Engineering [31, 32, 33, 34].

The application of GT is a rigorous research method enabling systematic and evidence-based development of theory, where data collection and analysis allows the researcher's emerging analysis to shape data collection procedures [35, 29]. Steps in Grounded Theory has several activities such as coding, memo writing, theoretical sampling, theoretical classification and integration [31, 32, 33, 34].

Stages of a Qualitative Research Process Five iterative and incremental stages are present in a qualitative research process [14]: Φ_1 *Problem statement*, Φ_2 *Data acquisition*, Φ_3 *Data management*, Φ_4 *Analysis* and Φ_5 *Report* [10, 14]. *Problem statement* Φ_1 refers to the theory review stage, formulation of research questions, the definition of methodologies, and construction of the theoretical framework. Data acquisition Φ_2 is the phase devoted to data collection, exploration, cleaning, and reliability verification. *Data management* Φ_3 refers to metadata generation, evaluation, and contextualization. *Analysis* Φ_4 consists of a round of experiments and measurements, incorporating debates and reflections on the results of previous stages. *Reporting* Φ_5 is the final phase devoted to visualization, evaluation, writing process, and final publication of scientific work.

Model of a Qualitative Research Process Workflow Throughout the research process phases Φ (Φ_1 *Problem statement*, Φ_2 *Data acquisition*, Φ_3 *Data management*, Φ_4 *Analysis* and Φ_5 *Report*) an artefact α evolves. It is coded, categorised, and analysed to create a knowledge artefact in the phases. GT and qualitative research methodologies, in general, are non-linear. In this context, artefacts evolve from one stage to another and eventually remain in a loop for a more detailed analysis.

The following structure details the associated metadata of an artefact α as it evolves through iterative research phases Φ.

$$\alpha : \quad \begin{array}{ll} \langle & \texttt{ID}: \text{UInt}, \\ & \texttt{Description}: \texttt{Text}, \\ & \texttt{Source}: \text{Url}, \\ & \texttt{CreatorORCID}: \text{OrcidID}, \\ & \texttt{CreationTimestamp}: \text{UInt} \rangle \end{array}$$

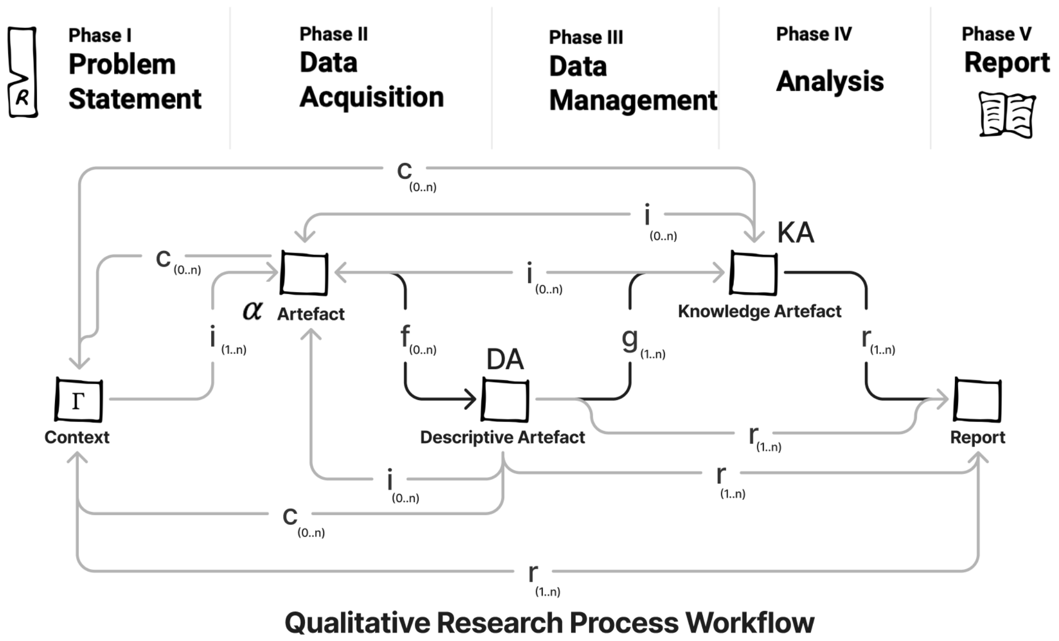

Figure 1: Model of Qualitative Research Process Workflow

`ID` represents a unique identifier of the artefact, `Description`, describes the artefact characteristics, `Source`, references the corresponding URL location , `CreatorORCID` refers to researcher identifier in ORCID format, and `CreationTimestamp` (timestamped of artefact creation in `UNIX` format).

Γ specifies the context of the research with the following structure: Theoretical framework description `TheoreticalFramework` (Grounded Theory), research questions `ResearchQuestions`, hypothesis `Hypotheses` and methods `Methods` (Interview, Cases Studies, Survey) applied in the research project.

$$\Gamma : \quad \begin{array}{ll} \langle & \texttt{TheoreticalFramework}\text{: Text,} \\ & \texttt{Methods}\text{: [Text],} \\ & \texttt{ResearchQuestions}\text{: [Text],} \\ & \texttt{Hypothesis}\text{: [Text],} \\ & \texttt{Findings}\text{: [Text]} \rangle \end{array}$$

The workflow model is shown in Figure 1, which represents the qualitative research process workflow model. We consider the context of the research Γ as an input of every operation in the model. The notation established in the figure can be described as follows[9]:

Researchers initially inquire about the research context Γ, and from the operator i, the raw artefact α is obtained. Subsequently, applying the f operator to the artefact α and considering the context Γ, researchers get a descriptive artefact (DA). Consecutively, applying a deep

[9] The cardinality of the operators is indicated by 0..n or 1..n

interpretation s of analytical categories and considering the context Γ, applying g, the Knowledge artefact (KA) is obtained. Different back and forth are present in qualitative research. For example, while analyzing KA, it's possible to reformulate the project's hypotheses and research questions. In this scenario, the context operator c applies. Finally, after several loops of interactions between data and analysis, the research team reached a consensus about research findings, writing the scientific report applying the operator r.

4. Use Case Scenario

This proposal emerges from an experimental study of an e-social science research project within a qualitative context. The project is called MENTOR (seMantic Exploration and curaTion of Open Hybrid Research). Workshops, interviews, and participant observation were employed better to understand the Grounded Theory (GT) research inquiry.

Several research questions are posed: What are the transformations, resignifications, continuities, and discontinuities between traditional and digital study practices? The research methodology was Grounded Theory, and the primary research method was the self-administered interview. All these attributes are part of the metadata that we consider part of the research context Γ and established in the first Problem Statement Φ_1 stage.

Initially, α represents a 'raw' artefact transcript of the self-administered interview obtained in the Data Acquisition Φ_2. The α artefact goes through a coding process f in the Data Management phase Φ_3. Each researcher considers a set of labels τ that arise from interpreting the artefact. This coding activity is a core part of the GT theoretical framework. Examples of tags are the following: τ_1 (paper vs. digital study practices), τ_2 (extracurricular readings), and τ_3 (aspects of virtual and in-person socialization). These tags are recorded as generating a new Descriptive artefact (DA). The analysis process begins once the researchers, in consensus, identify the descriptive categories to an adequate and rigorous descriptive interpretation of the phenomena. In this Φ_4 Analysis phase, researchers give meaning and semantics to the coding of the previous stage. At this stage, the operator g creates the Knowledge Artefact (KA). This KA is not a mere description but a deep analysis of findings regarding coding and the research context. An example of the application $g(f(\alpha), \Gamma)$ allows researchers to make findings such as "changes in study practices, skills for reading comprehension and ability to synthesize remain closely linked to the University academic journey." Once the research team agrees upon this analysis, the result is published in the scientific report at stage Φ_5.

Linearity does not apply in qualitative methodologies, especially in GT. It is the iteration and spiralling back and forth that gives methodological rigor. Thus, it is possible to remain in a loop in a specific stage or eventually return to the first stage. Therefore, in each stage, we conceptualise a back and forth as i as the operation to inquire and c as the corresponding operator for changing the research context. The inquiry i involves reanalyzing without changing the context Γ. On the other hand, the context c implies a significant change Γ, for example, a change in the objectives, methods, hypotheses, or research questions. This iterative back-and-forth process, in all operations, is what we call a Researcher-in-the-Loop intervention that establishes the criteria for progress and reflection regarding the object of study.

Our research represents a first attempt to model the artefact's lifecycle of highly diverse content generated on several stages of a qualitative research process. The presented workflow represents the lifecycle of artefacts and operators as they evolve throughout the phases of the qualitative research process, such as Grounded Theory. From this work, new research questions arise: How do researchers intervene in the process? How is research data curation redefined? Which qualitative methodologies, other than Grounded Theory, follow this model or establish a different model of research data curation?

5. Conclusion and Future Work

This paper presents a model to describe the life cycle of artefacts along the workflow of a qualitative research process. This proposal emerges from an experimental study of an e-social science research project in a qualitative research context. The project is called MENTOR (seMantic Exploration aNd curaTion of Open hybrid Research). To validate our proposal, we present a use-case scenario of the experimental outcomes of the research project.

Our research is the first attempt to model the curation process of highly diverse content generated by qualitative research. We argue that modeling the life cycle of content is a relevant part of the curation process. A model of the lifecycle of knowledge artefacts within a qualitative research process can provide insight into the reliability and transparency of the process itself. In future work, we plan to model consensus protocols adopted by scientists to validate content and decide the content evolution of the qualitative research process.

References

[1] P. Li, Knowledge and meta-knowledge: from the generating of knowledge to the management of knowledge, in: 2018 International Conference on Management and Education, Humanities and Social Sciences (MEHSS 2018), Atlantis Press, 2018, pp. 73–79.

[2] D. Garkov, C. Müller, M. Braun, D. Weiskopf, F. Schreiber, research data curation in visualization: Position paper, in: 2022 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV), IEEE, 2022, pp. 56–65.

[3] G. Kemp, CURARE: curating and managing big data collections on the cloud, Ph.D. thesis, Université de Lyon, 2018.

[4] G. Vargas-Solar, G. Kemp, I. Hernández-Gallegos, J. A. Espinosa-Oviedo, C. F. Da Silva, P. Ghodous, Demonstrating data collections curation and exploration with curare, in: EDBT/ICDT Conference 2019, 2019, p. 4.

[5] G. Vargas-Solar, G. Kemp, I. Hernández-Gallegos, J. A. Espinosa-Oviedo, C. F. Da Silva, P. Ghodous, Exploring and curating data collections with curare, in: Proceeding of the 35eme Conférence sur la Gestion de Données–Principes, Technologies et Applications, 2019.

[6] J. Leipzig, D. Nüst, C. T. Hoyt, K. Ram, J. Greenberg, The role of metadata in reproducible computational research, Patterns 2 (2021) 100322.

[7] L. M. Given, The Sage encyclopedia of qualitative research methods, Sage publications, 2008.

[8] K. A. d. G. e Silva, A. P. Costa, Study on contexts and stages of digital content curation models: Guidelines for use in qualitative analysis software, The Qualitative Report 28 (2023) 2980–2994.

[9] Y. Minamiyama, H. Takeda, M. Hayashi, M. Asaoka, K. Yamaji, A study on formalizing the knowledge of data curation activities across different fields, Plos one 19 (2024) e0301772.

[10] A. Adorjan, Towards a researcher-in-the-loop driven curation approach for quantitative and qualitative research methods, in: European Conference on Advances in Databases and Information Systems, Springer, 2023, pp. 647–655.

[11] S. Karcher, D. Kirilova, C. Pagé, N. Weber, How data curation enables epistemically responsible reuse of qualitative data, The Qualitative Report (2021).

[12] G. Demartini, J. Yang, S. Sadiq, Report on the 1st workshop on human-in-the-loop data curation (hil-dc 2022) at cikm 2022, in: ACM SIGIR Forum, volume 56, ACM New York, NY, USA, 2023, pp. 1–8.

[13] M. Vuorre, J. P. Curley, Curating research assets: A tutorial on the git version control system, Advances in Methods and Practices in Psychological Science 1 (2018) 219–236.

[14] A. Adorjan, G. Vargas-Solar, R. Motz, Towards a human-in-the-loop curation: A qualitative perspective, in: 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), IEEE, 2022, pp. 1–8.

[15] A. Zuiderwijk, R. Shinde, W. Jeng, What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption, PloS One 15 (2020).

[16] G. Vargas-Solar, J. Darmont, A. Adorjan, J. A. Espinosa-Oviedo, C. Hara, S. Loudcher, R. Motz, M. Musicante, J.-L. Zechinelli-Martini, Dataversifying natural sciences: Pioneering a data lake architecture for curated data-centric experiments in life\& earth sciences, arXiv preprint arXiv:2403.20063 (2024).

[17] A. Trisovic, M. K. Lau, T. Pasquier, M. Crosas, A large-scale study on research code quality and execution, Scientific Data 9 (2022) 60.

[18] C. Boyd, Use of optional data curation features by users of harvard dataverse repository, Journal of eScience Librarianship 10 (2021).

[19] E. Coburn, L. Johnston, Testing our assumptions: preliminary results from the data curation network, Journal of eScience Librarianship 9 (2020).

[20] S. Choudhury, C. Huang, C. L. Palmer, Updating the dcc curation lifecycle model, International Journal of Digital Curation 15 (2020) 12–12.

[21] S. Higgins, The dcc curation lifecycle model, International journal of digital curation 3 (2008) 134–140.

[22] H. Lee, S. Yoon, Z. Park, "semantic" in a digital curation model, Journal of Data and Information Science 5 (2020) 81–92.

[23] H. L. Rhee, A new lifecycle model enabling optimal digital curation, Journal of librarianship and information science 56 (2024) 241–266.

[24] L. R. Johnston, J. Carlson, C. Hudson-Vitale, H. Imker, W. Kozlowski, R. Olendorf, C. Stewart, M. Blake, J. Herndon, T. M. McGeary, et al., Data curation network: A cross-institutional staffing model for curating research data, International Journal of Digital Curation 13 (2018) 125–140.

[25] C. Simone, Knowledge artifacts: the implications of incommensurable dimensions for

their design, Data technologies and applications 52 (2018) 130–147.

[26] F. Cabitza, A. Cerroni, C. Simone, Knowledge artifacts within knowing communities to foster collective knowledge, in: Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, 2014, pp. 391–394.

[27] B. G. Glaser, A. L. Strauss, E. Strutzel, The discovery of grounded theory; strategies for qualitative research, Nursing research 17 (1968) 364.

[28] A. Strauss, J. Corbin, Grounded theory methodology: An overview. (1994).

[29] K. Charmaz, Constructing grounded theory: A practical guide through qualitative analysis, sage, 2006.

[30] B. G. Glaser, Conceptualization: On theory and theorizing using grounded theory, International journal of qualitative methods 1 (2002) 23–38.

[31] C. B. Seaman, Qualitative methods in empirical studies of software engineering, IEEE Transactions on software engineering 25 (1999) 557–572.

[32] R. Hoda, Socio-technical grounded theory for software engineering, IEEE Transactions on Software Engineering 48 (2021) 3808–3832.

[33] R. Hoda, Technical briefing on socio-technical grounded theory for qualitative data analysis, in: Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings, 2024, pp. 436–437.

[34] R. Hoda, Decoding grounded theory for software engineering, in: 2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), IEEE, 2021, pp. 326–327.

[35] K. Charmaz, The legacy of anselm strauss in constructivist grounded theory, in: Studies in symbolic interaction, volume 32, Emerald Group Publishing Limited, 2008, pp. 127–141.