FORS GUIDES

to survey methods & data management

Understanding Documentation and Metadata: A Guide for Social Scientists

Auriane Marmier[1]

[1] FORS – Swiss Centre of Expertise in the Social Sciences

FORS Guide No. 27, Version 1.0

September 2026

Abstract:

This guide explores the role of documentation and metadata in the management of social science research data. It traces the historical development of documentation and metadata, clarifies their relationship, key differences, and proposes a layered approach including project, file, data and repository levels. Finally, the guide provides practical recommendations to support transparent, reusable, well-documented research data.

Keywords: Research data management, documentation, metadata, social science.

How to cite:

Marmier, A. (2026). Understanding Documentation and Metadata: A Guide for Social Scientists. *FORS Guides*, 27, Version 1.0, 1-29. https://doi:10.24449/FG-2026-00027

The FORS Guides to survey methods and data management

The FORS Guides offer support to researchers and students in the social sciences who intend to collect data, as well as to teachers at university level who want to teach their students the basics of survey methods and data management. Written by experts from inside and outside of FORS, the FORS Guides are descriptive papers that summarise practical knowledge concerning survey methods and data management. They give a general overview without claiming to be exhaustive. Considering the Swiss context, the FORS Guides can be especially helpful for researchers working in Switzerland or with Swiss data.

FORS, Géopolis, CH-1015 Lausanne
www.forscenter.ch/publications/fors-guides
Contact: info@forscenter.ch

Statement on the use of artificial intelligence:

For this guide, the author used the model Sonnet 5 of Claude AI for proofreading, rephrasing, and some literature research. After using this tool/service, the author has carefully reviewed and edited the content as necessary and takes full responsibility for the final publication.

1. INTRODUCTION

Over the past two decades, the rise of the Open Science movement has profoundly reshaped how research is produced, managed, and shared. By promoting transparency, accessibility, and collaboration — through practices such as open access publishing, open peer review, and open data sharing — Open Science has placed new expectations on researchers across all disciplines (Bartling & Friesike, 2014; Berkowitz & Delacour, 2022). What began as a bottom-up movement driven by researchers and scholarly communities has progressively been taken up by governments, funding bodies, and international organisations. They now see open practices as serving both scientific and civic goals: accelerating discovery, strengthening the credibility of research, and ensuring that publicly funded knowledge remains accessible and useful to society at large (Vicente-Saez & Martinez-Fuentes, 2018). This shift reflects a growing recognition that science is most valuable when it is transparent, verifiable, and built on foundations that others can understand and extend. Given that, these convictions have been translated into concrete policy obligations. National research agencies and international funders such as the Swiss National Science Foundation (SNSF) and Horizon Europe now mandate Open Research Data (ORD) practices, requiring researchers to make their data available, well-described, and reusable wherever possible (European Commission, n.d.; Swiss National Science Foundation, n.d.).

Responding to these demands requires more than simply "copy-paste" data in an online repository. It requires a systematic approach to research data management (RDM). RDM is described as the set of practices and processes through which data are collected, organised, documented, stored, protected, and shared across the research lifecycle (Corti et al., 2014; Whyte & Tedds, 2011). Good RDM ensures that data remain usable and shareable not only during an active project, but also long after it has ended. This concerns the original research team, reviewers, and future researchers who want to understand or reuse the data. By contrast, poor data management can lead to data loss and duplication of effort, eroding scientific value. Consequently, datasets that cannot be understood cannot be trusted, and datasets that cannot be found cannot be used (Borgman, 2017; Tenopir et al., 2011).

To guide researchers and institutions in this RDM effort, the FAIR principles have emerged as a compass for good data management. Developed by Wilkinson et al. (2016), these principles provide guidelines for enhancing the Findability (discoverability), Accessibility, Interoperability and Reusability of digital resources. They place strong emphasis on enabling machines to automatically process data, thereby placing documentation and metadata at the heart of responsible data practice. Documentation can be broadly defined as information about a research object describing its content, context, and conditions of production, while metadata can be considered a structured subcategory of documentation, standardised to enable the discovery and exchange of the data[1]. For data to be findable, they must carry detailed descriptive metadata and be registered in repositories that support search and discovery. For data to be accessible, they must be retrievable through open, standardised protocols, and crucially, descriptive metadata should remain available to users even after the underlying data have been removed or restricted. For data to be interoperable, metadata must use standardised vocabularies and formats that allow datasets to be combined and compared across studies and systems. And for data to be reusable, metadata must include clear information about provenance, licensing, and context, so that future users can assess whether the data are fit for their

[1] Both concepts will be defined more precisely in the following sections.

purposes (Wilkinson et al., 2016). The FAIR principles make clear that RDM is inseparable from documentation: without adequate descriptive information, data can be neither found nor reused (Borgman, 2017; Wilkinson et al., 2016).

This is particularly the case in the social sciences. Research data — understood broadly as any material collected or produced during a study, from survey datasets and interview transcripts to field notes and administrative records — are rarely self-explanatory. Their meaning is shaped by the conditions under which they were collected, the theoretical frameworks guiding the research, and the methodological choices made by the researchers (Corti et al., 2014). A simple illustration makes this concrete: in Douglas Adams's *The Hitchhiker's Guide to the Galaxy*, the number 42 is famously the answer to the Ultimate Question of Life, the Universe, and Everything, but because no one knows what the question is, the answer is completely meaningless (Adams, 1979). Stripped of its context, 42 is just an arbitrary figure. In the same way, data without documentation are just uninterpretable. Documentation and metadata act as a bridge between raw data and meaningful analysis, transforming isolated observations into interpretable and reusable knowledge (Borgman, 2017). Without them, data may be technically accessible yet practically unusable (Borgman, 2017; Wilkinson et al., 2016).

Therefore, this guide aims to explain and describe the various aspects of documentation and metadata and demonstrate “how to do it”. To this end, the following part of this guide clarifies the concepts of documentation and metadata for social scientists by outlining their historical development. The third part describes the layered approach generally used to provide information on a research project, highlighting the complementary roles of documentation and metadata throughout the research lifecycle. Here, researchers can access practical tools and conceptual frameworks for making their data more transparent, findable, and reusable. Finally, the guide concludes with key recommendations for social scientists who wish to embrace the principles of open science and the FAIR movement from the perspective of documentation and metadata.

2. DOCUMENTATION AND METADATA

2.1 HISTORICAL PERSPECTIVE

Yet despite the central role of documentation and metadata in contemporary research practice, the boundary between documentation and metadata remains poorly defined in the social sciences. In both practice and RDM training materials, the two terms are frequently used interchangeably (Edwards et al., 2011; Niu, 2009; Niu & Hedstrom, 2008) and their meanings vary across contexts, research traditions, and individual preferences. Some definitions of metadata emphasise technical characteristics such as file formats or metadata standards (Riley, 2017; Woodley, 2008), while others focus on descriptive content such as methodological information or variable definitions (Humphrey, 2014). This conceptual inconsistency complicates the standardisation of practices and limits the potential for data reuse and interoperability (Niu & Hedstrom, 2008; Tenopir et al., 2011; Wilkinson et al., 2016).

Thus, to understand why the terms 'metadata' and 'documentation' are often used interchangeably and are sometimes difficult to distinguish, this guide traces the historical development of descriptive practices. From pre-digital documentation traditions to contemporary metadata standards, this guide covers three major eras: the paper era, the digital era and the web era.

The Paper Era: Early Documentation Practices

Efforts to describe and organise information are deeply rooted in human history, and early systematic documentation practices can be traced to ancient libraries. Historical accounts suggest that Zenodotus, the first librarian of the Library of Alexandria (third century BCE), attached small tags to scrolls containing information such as author, title, and subject. According to the legend, this allowed him efficient retrieval without unrolling the material (Bellier-Chaussonnier, 2002; Lerner, 2001). The principle of attaching standardised descriptive information to a physical object to enable its retrieval proved remarkably durable. The first French card catalogue in 1791, developed to inventory millions of books and manuscripts confiscated from aristocratic and clerical libraries following the French Revolution, applied the same logic at an unprecedented scale: one card per item, recording author, title, format, and location (Guilleminot-Chrétien, 1989; Lerner, 2001). Building on this same drive towards standardisation, Melvil Dewey's Decimal Classification in 1876 organised books by subject through a hierarchical numerical system (Mitra, 2016). In library and information science, these practices are recognised as foundational forms of documentation — systematic efforts to describe objects in ways that enable their retrieval and reuse (Buckland, 1991; Bowker, 2001).

When empirical social science expanded rapidly in the post-war decades, this same descriptive logic — attaching standardised information to objects to enable their retrieval and interpretation — was applied to research materials. Fostered by post-war social surveys, census operations, opinion polling, and early longitudinal studies, the volume of data requiring systematic description grew substantially between the 1940s and the 1980s (Hauptmann, 2024; Inter-University Consortium for Political and Social Research [ICPSR], 2012; Shankar et al., 2016). Researchers relied primarily on questionnaires and interviews, methodological reports, field notes, and manually created codebooks to make large surveys and longitudinal datasets usable. These documents served as crucial interpretive tools, offering details on sampling strategies, question wording, coding decisions, interviewer instructions, and the wider sociopolitical context of data collection (Shankar et al., 2016).

However, documentation practices during this period were highly heterogeneous. Each research team or institution developed its own conventions, which varied widely across disciplines such as sociology, political science, education, and economics (Shankar et al., 2016). The consequences of this fragmentation were tangible: a codebook produced by one research team might use different variable naming conventions, missing value codes, or category labels than those of another, making it difficult or impossible to comprehend and combine datasets across studies. Much of the methodological knowledge underpinning these datasets remained tacit, embedded in local procedures or scattered across notebooks, memos, and coding sheets (Niu, 2009; Niu & Hedstrom, 2008). When researchers moved on or retired, that knowledge often disappeared with them. Consequently, the ability to understand or reuse data outside the originating project was often limited, highlighting the need for more systematic and shared approaches to documentation (Borgman, 2017; Shankar et al., 2016).

The Digital Era: Standardisation, Archives, and Early Metadata

This rapid accumulation of research materials and the fragmentation of documentation practices have posed significant preservation and documentation challenges and have led to the creation of dedicated data archives. The Roper Centre for Public Opinion Research, founded in 1947 at the University of Connecticut, was among the first institutions to systematically archive and preserve public opinion and survey data, making them available for secondary

analysis (Hauptmann, 2024). Shortly after, the Inter-University Consortium for Political and Social Research (ICPSR), founded in 1962 at the University of Michigan, went further by developing a collaborative model in which member institutions could both deposit and access social science datasets. ICPSR rapidly became an international reference for data archiving and developed some of the earliest codebook standards for social science data (Rasmussen, 2014; Shankar et al., 2016). Together, these institutions sought to centralise research materials, formalise documentation practices, and promote more systematic approaches to data description — laying the groundwork for what would later become standardised social science data documentation (Rasmussen, 2014; Vardigan et al., 2008).

The arrival of computers fundamentally reshaped these documentation practices by introducing machine-readable descriptions. A major milestone was the creation of the Machine-Readable Cataloguing (MARC) format by the Library of Congress in the 1960s (McCallum, 2002a, 2002b; Zeng & Qin, 2020). MARC converted traditional catalogue descriptions (i.e. previously recorded on paper cards) into standardised, machine-readable formats that could be processed, searched, and shared electronically across library systems. In a standardised and widely adopted form for the first time, descriptive information about an object could be not only read by humans but also interpreted and exchanged by computers (McCallum, 2002a; Zeng & Qin, 2020). Building on this principle, the 1990s saw the emergence of the Data Documentation Initiative (DDI), an international standard designed specifically to describe social, behavioural, and economic research data (Vardigan et al., 2008). Unlike earlier documentation formats, DDI explicitly linked social science documentation with machine-readable metadata. This allows researchers to describe not only the basic bibliographic information of a dataset but also its internal structure, including variable names, question wording, response categories, and sampling procedures. DDI represented a conceptual shift by treating documentation not only as a narrative context but also as structured data that computers could process (Rasmussen, 2014; Vardigan et al., 2008). This development started to blur the traditional boundary between documentation and metadata.

The Web Era: Metadata Everywhere and Parallel Traditions

The emergence of the World Wide Web in the early 1990s marked a third major turning point in the history of descriptive practices. Where the digital era had introduced machine-readable descriptions, the web transformed the landscape entirely by creating an open, networked, and globally interconnected information environment. For the first time, information of extraordinary scale and diversity needed to be described, located, and exchanged across open, decentralised systems — making the ability to discover and share information a fundamental infrastructural challenge (Gilliland, 2008; Greenberg, 2003). Existing cataloguing technologies, such as MARC, were adapted to support online catalogues and metadata exchange (Arms, 2012), while new web-specific standards such as the Dublin Core Metadata Element Set extended metadata practices far beyond their library origins (Weibel et al., 1998). Metadata became essential not only for describing documents but also for enabling search engines, supporting persistent identifiers, facilitating digital preservation, and powering machine-actionable workflows, including AI-driven retrieval (Wilkinson et al., 2016; Yang et al., 2025).

Among the standards that emerged to meet the demands of this open environment, Dublin Core is perhaps the most illustrative example. Dublin Core was developed in 1995 at a workshop in Dublin, Ohio, by an international group of librarians, computer scientists, and researchers. It proposed a simple, interoperable set of fifteen descriptive elements — such as title, creator, subject, and date — that could be applied to any digital resource, regardless of

discipline or format (Dublin Core Metadata Initiative, 2012; Weibel et al., 1998). Its simplicity and flexibility made it one of the most widely adopted metadata standards on the web, and it remains in common use today in generalist data repositories.

Social science data archives were not passive observers of these developments. From the late 1990s, established institutions such as ICPSR and the emerging network of European archives coordinated through the Consortium of European Social Science Data Archives (CESSDA) progressively migrated their holdings to web-based platforms, making datasets searchable and accessible online for the first time (Vardigan et al., 2008). This transition created new pressures: for data to be discoverable through web-based search systems, archives needed standardised, machine-readable metadata that could be harvested and indexed across institutions. The DDI standard, already developed for social science data description, was progressively extended and adopted across member archives as a common metadata language, enabling for example a researcher in Switzerland to search and retrieve a dataset deposited in the United Kingdom using the same descriptive categories (Rasmussen, 2014; Vardigan et al., 2008). More recently, national and institutional repositories such as SWISSUbase have built on these foundations, combining DDI-compliant metadata for repository discovery with richer project-level documentation for contextual understanding (SWISSUbase, n.d.).

Meanwhile, social science researchers continued to produce narrative documentation in line with long-standing academic traditions — README files, methodological reports, codebooks, and field notes remained central tools for describing the context, content, and conditions of data collection (Corti et al., 2014; Shankar et al., 2016). They evolved incrementally alongside the new metadata standards emerging from library and web-based infrastructures, rather than being replaced by them. As a result, the web era consolidated the coexistence of two parallel descriptive traditions: on one side, documentation i.e. human-oriented, narrative, flexible, and project-specific, rooted in the practices of academic research (Corti et al., 2014); on the other side, metadata i.e. machine-readable, standardised, and repository-driven, shaped by library, archival, and web-based infrastructures (Riley, 2017; Zeng & Qin, 2020).

Understanding this historical divergence helps explain why defining these two concepts precisely — and understanding how they complement each other — remains both challenging and necessary for contemporary social science researchers.

2.2 CONCEPTUAL DEFINITIONS

Although these two traditions emerged from different contexts and served different primary purposes, they are based on the same idea of providing information. For example, a codebook can be structured as machine-readable metadata, while a DDI record can contain rich narrative descriptions of methodology. Consequently, the boundary between documentation and metadata is often blurred and contested in both practice and literature (Edwards et al., 2011; Niu, 2009; Niu & Hedstrom, 2008). Therefore, based on the historical overview presented above and the existing literature (Corti et al., 2014; Humphrey, 2014; Niu, 2009; Niu & Hedstrom, 2008; Van den Eynden & Chatsiou, 2011), the following definitions of documentation and metadata are adopted for this guide.

Documentation – The Rosetta Stone

Documentation refers to a broad set of materials that provide narrative, contextual, and procedural information about research projects or datasets (Borgman, 2017; Niu & Hedstrom, 2008). It includes any information that helps human readers understand how the data were produced,

under what conditions, and how they should be interpreted. Documentation may be structured or unstructured, narrative or standardised, and it may operate at different levels of the research process, e.g. at the level of the project, the files and folders, or the data.

Common forms of documentation in the social sciences include codebooks and questionnaires, interview guides and field notes, methodological reports, data collection protocols, research diaries or memos, and consent forms or ethics documentation, among other project-specific materials. Documentation is therefore primarily a human-oriented resource, designed to preserve the meaning, clarity, and interpretability of a project and its data (Corti et al., 2014; Van den Eynden & Chatsiou, 2011).

Metadata – The Machine Handshake

Metadata is a structured and standardised subset of documentation. Where documentation is designed to be read and interpreted by human users, metadata presents information in a format that can be processed, indexed, and exchanged by digital systems (Edwards et al., 2011; Riley, 2017; Vardigan et al., 2008; Zeng & Qin, 2020). Its primary function is to provide systematic, precise descriptions that enable discovery, interoperability, and automated processing within digital infrastructures. Following the established metadata literature (Riley, 2017; Zeng & Qin, 2020), three main categories of metadata can be distinguished:

- **Descriptive metadata** helps users find and identify a resource by capturing key information such as the author's name, the title, the date of creation, and relevant keywords. For digital objects, one of the most widely used forms of descriptive metadata is a persistent identifier (PID), such as a Digital Object Identifier (DOI), which provides a stable reference that facilitates discovery and citation) (Riley, 2017; Zeng & Qin, 2020).
- **Administrative metadata** records the management and technical information needed to maintain and govern a resource over time, including its format, rights status, and preservation history:
    - ***Technical metadata*** documents the technical characteristics of a digital file, such as its format, size, and the specifications needed to open, display, or process it. This is essential for ensuring that data can be accessed and reused with appropriate tools.
    - ***Preservation metadata*** captures the information required to maintain a resource over time, such as version history, data processing history, or migration actions. In the social sciences, preservation metadata is particularly important for longitudinal projects involving multiple data collection periods. For instance, version history enables researchers to better understand how the study has evolved.
    - ***Rights metadata*** describes the intellectual property and usage conditions of a resource, including ownership, licensing, and permissions. For social science data, this is particularly important when working with sensitive materials, restricted-access datasets, or third-party sources (Riley, 2017; Zeng & Qin, 2020).
- **Structural metadata** captures the internal organisation of a resource, specifying how its component parts relate to one another e.g. how files, variables, or dataset waves are connected. Structural metadata helps researchers understand how different parts of a social-science dataset fit together and how they should be used (Banerjee, 2025; Riley, 2017).

2.3 KEY DIFFERENCES BETWEEN DOCUMENTATION AND METADATA

While documentation and metadata both serve to describe research materials, they do so in different formats, target different audiences, and fulfil different functions within digital research infrastructures. Understanding these differences is essential for researchers seeking to make their data transparent, discoverable, and reusable.

Format and Medium of Documentation and Metadata

Documentation is generally stored in a separate file, created by researchers (e.g., README files, codebooks or methodological reports) and annexed to a dataset when researchers deposit it (Cornell Data Services, n.d. a; Corti et al., 2014). Yet, documentation can also be embedded directly within data files, as variable and value labels in a quantitative dataset or as header information in a qualitative transcript. Metadata, by contrast, is typically entered into the web interface of a data repository, which structures it according to specific standards (e.g., Dublin Core or DDI) (Riley, 2017; Zeng & Qin, 2020).

Audience and Purpose of Documentation and Metadata

Documentation speaks primarily to human readers, supporting understanding, interpretation, and reuse by providing narrative context and methodological detail (Van den Eynden & Chatsiou, 2011). By contrast, metadata speaks primarily to machines and information systems, supporting discovery, citation, automated processing, and interoperability across digital infrastructures (Niu, 2009; Niu & Hedstrom, 2008). That said, the two are not opposed: good metadata makes data findable (Borgman, 2017), while good documentation makes it usable, and both are necessary for meaningful data sharing (Zeng & Qin, 2020).

Documentation and Metadata Structure

In general, documentation is flexible and narrative, and its structure varies according to disciplinary conventions and the nature of the research methods employed (Humphrey, 2014). Metadata, by contrast, is highly structured and standardised, following standards such as the Dublin Core or DDI (Riley, 2017).

Key differences between the two terms are summarised in Figure 1 below:

Figure 1. Key differences between documentation and metadata

Aspect	Metadata	Documentation
Format	Usually entered into the web interface of a data repository	Stored in a separate file and annexed to a data repository
Creator	Generated by standards or data repositories and completed by researchers	Researchers
Audience	Machines and information systems	Human readers
Purpose	Discovery, interoperability, citation, automated processing	Interpretation, contextualisation, reuse
Structure	Structured, standardised	Narrative, flexible
Examples	Metadata fields in data repositories	Codebooks, field notes, protocols, etc.

3. METADATA & DOCUMENTATION: TOWARDS A LAYERED APPROACH

While documentation and metadata serve distinct but complementary purposes, they also operate at different levels of granularity within a research project. No single file or record can capture all the information needed to make projects or data understandable and reusable — descriptive information must instead be provided across several levels of the research process (Corti et al., 2014; UK Data Archive, n.d. a). We distinguish four levels: (1) the research project, (2) files and folders, (3) the data themselves, and (4) the deposit. Together, these layers form a coherent descriptive system that supports both human interpretation and machine-readable reuse.

3.1 THE RESEARCH PROJECT LEVEL — THE FRONT DOOR OF A PROJECT

Project-level documentation refers to the set of materials that provide an overarching description of a research project, i.e. its aims, design, context, and conditions of production. It is the first and broadest layer of the documentation system, and it is essential for ensuring that anyone encountering the data — whether a colleague joining the project midway, a reviewer assessing a submission, or a researcher discovering the dataset years later — can understand what the study was about, how it was conducted, and how the data it produced should be interpreted (CESSDA Training Team, 2020, p. 47; Finnish Social Science Data Archive, n.d. b; UK Data Archive, n.d. c).

Project-level documentation seeks to answer a core set of questions that are widely recognised across data management frameworks (CESSDA Training Team, 2020; Finnish Social Science Data Archive, n.d. b; UK Data Archive, n.d. c): *What is the study about? Who conducted the study? When and where was it conducted? Why was it conducted? How was it conducted? What data were collected? How are the data structured? What ethical/legal constraints apply?*

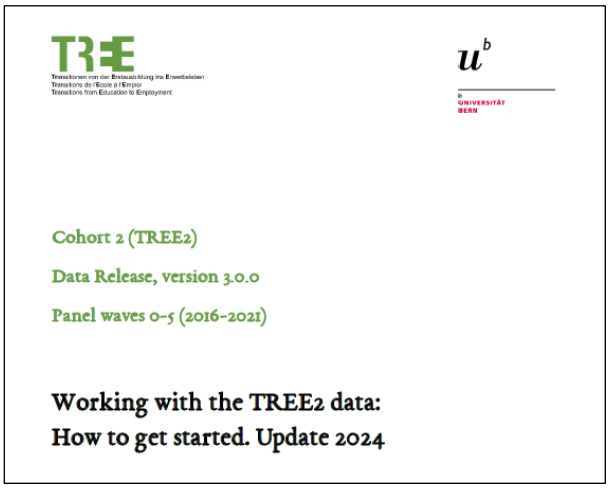
To ensure completeness and consistency, project-level documentation typically includes the following elements (CESSDA Training Team, 2020; Finnish Social Science Data Archive, n.d. b; UK Data Archive, n.d. c).

- **Purpose, scope, and coverage**: A summary of what the study aims to investigate, the population studied, and the temporal and geographic coverage
- **Research context and design**: A description of the theoretical framework, research questions, and methodological approach.
- **Data collection methods**: Information about instruments used (e.g., surveys, interviews), sampling strategies, and fieldwork procedures.
- **Data preparation and transformation**: Notes on cleaning, coding, anonymisation, and any manipulations applied to the raw data.
- **Cataloguing and citation information**: Persistent identifiers (e.g., DOIs), dataset titles, and recommended citation formats.
- **Contact and organisational details**: Names and affiliations of principal investigators, research teams, and institutions.
- **Funding and ethical approvals**: Grant numbers, funders, and documentation of ethics review processes.

- **Links to related materials**: Questionnaires, consent forms, methodological reports, publications, and grant proposals.

Although there is no single standardised format for project-level documentation, it often takes one of the following forms: a README file, which provides a concise overview of the project and its materials (Cornell Data Services, n.d. a), a user guide, which offers more detailed guidance on how to navigate and understand the project (van Zuilen et al., 2006), or a technical or methodological report, which documents the research process in depth, including sampling procedures and data collection instruments (Heers & Hupka-Brunner, 2023). The choice of format will depend on the project's complexity, the intended audience, and any requirements set by funders or data repositories.

Figure 2. Example of a ReadMe file



Note. From *Working with the TREE2 data: How to get started*, by TREE, 2024, SWISSUbase.

3.2 THE FILES AND FOLDERS LEVEL — THE PROJECT FLOOR PLAN

While project-level documentation provides the overarching context for a research study and can be seen as the front door of a research project, file and folder-level documentation operates at a more granular level and can be seen as the project floor plan. This level ensures that the structure, organisation, and contents of the research materials are understandable, navigable, and usable — both during the active phase of the project and long after its conclusion (CESSDA Training Team, 2020, pp. 43–46; Finnish Social Science Data Archive, n.d. a). This level of documentation is essential for maintaining data integrity, facilitating collaboration, and ensuring that materials can be preserved and reused over time (Corti et al., 2014).

This level of documentation is particularly important in two contexts: collaborative environments where multiple researchers interact with shared folders and need to locate, identify, and update files easily; and archival settings, where future users must be able to locate and interpret the materials independently. In both cases, file and folder documentation seeks to answer the following questions (CESSDA Training Team, 2020, pp. 43–46; Finnish Social Science Data Archive, n.d. a): *How are files organised and named? What does each file contain? How do files relate to each other? What software or formats are required to open and interpret them?*

Good file and folder documentation is based on three interconnected practices. The first is developing a logical and consistent folder hierarchy that reflects the structure of the research process and allows users to find materials easily (Finnish Social Science Data Archive, n.d. a;

Sherman Center for Digital Scholarship, 2025). The second is adopting a clear and descriptive file naming strategy that follows a standardised format — including information such as the date, version, and content of the file (CESSDA Training Team, 2020, pp. 43–46). The third is implementing a versioning strategy that tracks changes to files over time, either through naming conventions using suffixes (e.g., v1_0, v2_1) or through version control software such as Google Workspace, Microsoft SharePoint, or GitHub (Cornell Data Services, n.d. b; Humphrey, 2014).

While plain text files, Word documents, and spreadsheets can all be used to document file and folder organisation, a README file placed at the root of the directory is widely recognised as a simple yet powerful tool for this purpose (CESSDA Training Team, 2020, pp. 43–46; Cornell Data Services, n.d. a; Finnish Social Science Data Archive, n.d. a). A well-constructed README file for a research project typically includes a folder structure overview, the names and contents of files and folders, the file and folder naming conventions adopted, and the versioning strategy used. It should be updated regularly as the project evolves, and additional README files or descriptive notes can be added to subfolders or individual files where needed. The Cornell Data Services README template (Cornell Data Services, n.d. a) provides a useful starting point for researchers building this kind of documentation from scratch. Table 1 below provides an example of a list of files that could be included in a README file.

Table 1. Example of a list of files

File	**Comment**
chx_2016-2017_Data_v1.0.0.sav	Data file in SPSS format (men and women)
chx_2016-2017_Questionnaire_F.pdf	Questionnaire in French (original version)
chx_2016-2017_Questionnaire_D.pdf	Questionnaire in German (original version)
chx_2016-2017_Questionnaire_I.pdf	Questionnaire in Italian (original version)
chx_2016-2017_Questionnaire_E.pdf	Questionnaire in English (translated version)
chx_2016-2017_Book_F.pdf	Book in French (with summary in German and Italian)
chx_2016-2017_Summary_E.pdf	Summary in English
chx_2016-2017_TechnicalReport_E.pdf	Technical report (including "Instructions to experts", see appendix) in English
chx_2016-2017_Codebook_FDI.pdf	Codebook in French, German and Italian (men only, original version)
chx_2016-2017_Codebook_E.pdf	Codebook in English (men only, translated version)
chx_2016-2017_Syntax_Weighting.zip	Syntax for creating a weight for the male population to compare them to the female sample
chx_2016-2017_Syntax_Stay.zip	Syntax for creating a variable “stay” (“sejour”) as used in the ch-x book 2016/2017

Note. Adapted from *ch-x 2016/2017: Read me (list of files)*, by FORS, 2021, SWISSUbase.

This level of documentation is important because poor file organisation and inadequate documentation can reduce visibility, making it more difficult to locate the correct files. This can result

in unnecessary duplication of effort and, even worse, hinder reuse. Well-documented file systems, by contrast, enhance efficiency, reduce the risk of data loss, and support compliance with the FAIR principles. Once files are properly organised and documented, attention can turn to the content of the data themselves, i.e. the focus of the following section.

3.3 THE DATA LEVEL

Data-level documentation is the most granular layer of documentation activities. This documentation level goes further by providing detailed information about the structure, content and meaning of data. This level of documentation is essential for interpretation and reuse, as it ensures that both human users and machines can understand the data independently of the original research team (CESSDA Training Team, 2020, pp. 39–42; Finnish Social Science Data Archive, n.d. a, c, d).

In general, data-level documentation aims to answer the following questions: *What does each variable mean? What are the units, formats, and value ranges? How were the variables derived or transformed? What are the relationships between variables? What contextual information is needed to interpret the data?* (CESSDA Training Team, 2020, pp. 39–42; Finnish Social Science Data Archive, n.d. a, c, d).

Quantitative Data-Level

Quantitative data-level documentation provides the detailed and structured information necessary for understanding, interpreting and reusing numeric datasets. Its primary aim is to describe the variables, values and data structures within a dataset, ensuring that both human users and machines can understand the data.

In its embedded form, quantitative data-level documentation is stored directly within the data file i.e. as variable labels, value labels, and missing value codes defined within an SPSS, Stata, or R dataset. This approach has the advantage of keeping descriptive information close to the data themselves, reducing the risk that documentation and data become separated over time (Finnish Social Science Data Archive, n.d. d; UK Data Archive, n.d. d). In its non-embedded form, documentation is maintained in separate files such as codebooks, data dictionaries, or README files, which describe the dataset's structure and content independently of the data file (Finnish Social Science Data Archive, n.d. d; UK Data Archive, n.d. d).

Quantitative data-level documentation seeks to address questions such as (Finnish Social Science Data Archive, n.d. d; UK Data Archive, n.d. d): *What does each variable represent? What are the possible values and their meanings? How were variables derived or transformed? What are the units of measurement and data types? How should missing values be interpreted?*

To answer these questions, quantitative data-level documentation typically includes the following elements (Finnish Social Science Data Archive, n.d. d; UK Data Archive, n.d. d):

- Variable name (e.g., age, income)
- Variable label (e.g., “Age of respondent in years”)
- Value labels (e.g., 1 = Male, 2 = Female)
- Units of measurement (e.g., years, CHF, Likert scale)
- Data type (e.g., numeric, string, categorical)
- Missing value codes (e.g., -99 = Not applicable)
- Derivation notes (e.g., “Derived from Q3 and Q4”)
- Question text
- Response categories
- Skip patterns or filters (e.g., “Only asked if Q1 = Yes”)

Codebooks, one of the most common forms of non-embedded quantitative documentation, can be generated automatically using data analysis software such as SPSS, R and REDCap or created manually by researchers using templates (CESSDA Training Team, 2020, pp. 39–42). Table 2 provides an example of a codebook as a form of non-embedded quantitative data-level documentation (Tresch et al., 2019).

Table 2. Example of a codebook

File Selects2019_PES_v1.1.0							
#	**Name**	**Label**	**Type**	**Format**	**Valid**	**Invalid**	**Question**
1	userid	Personal ID	continuous	numeric-8.0	6664	0	-
2	imode	Mode of interview	discrete	numeric-1.0	6664	0	-
3	sample_g ..	Sample group	discrete	numeric-8.0	6664	0	-
4	intstart	Interview start date	discrete	character-20	5450	-	-
5	intend	Interview end date	discrete	character-20	5450	-	-
6	totaltime	Total interview length in minutes	continuous	numeric-8.2	5450	1214	-
7	finished	Completion of questionnaire	discrete	numeric-40.0	5450	1214	-
8	indate	Date of reception (paper questionnaire)	continuous	numeric-8.0	1214	5450	-
9	spr	Language of interview	discrete	numeric-8.0	6664	0	-
10	birthyear	Year of birth	discrete	numeric-8.0	6664	0	-
11	age	Age in years	discrete	numeric-8.0	6664	0	-
12	agecat	Age in categories	discrete	numeric-8.0	6664	0	-
13	sex	Gender	discrete	numeric-40.0	6664	0	-
14	f10000	Canton in which R has the right to vote in current national election (2019)	discrete	numeric-40.0	6664	0	-
15	f10100	Political interest	discrete	numeric-40.0	6645	19	-
16	f13400a	Attention to political news on TV	discrete	numeric-8.0	6508	156	-
17	f13400b	Attention to political news in newspapers (print or e-paper)	discrete	numeric-8.0	6311	353	-
18	f13400c	Attention to political news in free newspapers (print or e-paper)	discrete	numeric-8.0	6113	551	-

19	f13400d	Attention to political news on social media (e.g., Facebook, Twitter)	discrete	numeric-8.0	5994	670	-
20	f13400e	Attention to political news on online newssites (e.g., watson.ch, srf.ch, 20min.ch)	discrete	numeric-8.0	6089	575	-

Note. Adapted from *Post-Election Survey Codebook - EN*, by Selects, 2021, SWISSUbase.

Qualitative Data-Level

Qualitative data-level documentation provides the detailed information needed to contextualise, interpret and reuse non-numeric data such as interview transcripts, focus group recordings, field notes, diaries, open-ended survey responses, photographs, and videos, among other materials. (Finnish Social Science Data Archive, n.d. c; UK Data Archive, n.d. c). Unlike quantitative data, qualitative data are often unstructured and deeply linked to its production context, e.g. who collected them, under what circumstances, and with what intention. Documentation is therefore especially important for ensuring that qualitative materials remain comprehensible and usable beyond the original research context, and for allowing future users to assess their suitability for reuse, such as secondary analysis (Corti et al., 2014).

Qualitative data documentation can take both embedded and non-embedded forms. In its embedded form, it typically consists of contextual and descriptive information added directly to the material itself, e.g. a header at the beginning of an interview transcript recording the date, location, duration, and characteristics of the interview, or descriptive captions embedded within a photograph or video file (Finnish Social Science Data Archive, n.d. c; UK Data Archive, n.d. c). In its non-embedded form, it typically takes the shape of a structured list or register that compiles contextual information about an object, such as pictures, videos, or interviews, and includes details, including the names of the interviewer, interviewee's ID, the place and date where the interview took place, among others (Finnish Social Science Data Archive, n.d. c; UK Data Archive, n.d. c).

Qualitative data documentation seeks to address the following questions (Finnish Social Science Data Archive, n.d. c): *What is the context of the data collection? Who collected the data? Who are the participants, and what are their characteristics? What methods were used to collect, process and anonymise the data? Where was the data collected? When was the data collected?*

To address these questions, qualitative data-level documentation typically includes the following elements, which vary depending on the sensitivity of the data and the type of material being described (Finnish Social Science Data Archive, n.d. c):

- Interview: interview ID, age, gender, occupation, organisation, location, place of interview, date of interview, transcript file name, recording file name, etc
- Picture: photographer name, file name, description of photos, location of photos, date of photos, etc.
- Video: video director's name, file name, description of videos, location of videos, date of video, people in videos, time of videos, etc.

Figures 3 and 4 below are examples of qualitative data-level documentation.

Figure 3. Interview background information template — An embedded qualitative data-level documentation

Interview transcript

Interview code:
define and use a consistent and meaningful code for each interview
Interview date:
Interview duration:
Interview location:
Respondent(s):
define and use a consistent and meaningful code for each respondent for the entire research project (e.g. R1, R2 or for international research CHr1, CHr2, DEr3, etc.). If necessary and depending on the research, details and information can be added (e.g. age, profession, etc.) without compromising the confidentiality of the respondents
Interviewer:
define and use a consistent code for each of the interviewers for the entire research project, especially if the same person is doing multiple interviews (e.g. I1, I2, ...)

Start of interview

```
1  I : Font to be used for transcription: Courier New (Courier is recommended
2     for transcript analysis because it is a fixed-cast font, meaning that all
3     characters are exactly the same width)
4  R :
5  I :
6  R :
7
8    ...
9
```

End of interview

Note. From *Template: Interview transcript*, by FORS, n.d.

Figure 4. Photo list — A not-embedded qualitative data-level documentation

Political wall writings and stickers in Tampere and Helsinki 2018 - 2019

File name	Date	Photographer	Location	Description of the photo
Photo_01.jpg	13/05/2018	Matt Miller	Tampere city centre	Writing on underpass wall
Photo_01.jpg	13/05/2018	Matt Miller	Tampere University (main building)	Sticker on a lamp post
Photo_01.jpg	15/05/2018	Matt Miller	Pasila railway station	Sticker on a litter bin
Photo_01.jpg	16/05/2018	Matt Miller	Pasila railway station	Writing on station platform wall
Photo_01.jpg	15/02/2019	Matt Miller	University of Helsinki (Topelia)	Sticker on a park bench
Photo_01.jpg	22/02/2019	Matt Miller	Tampere railway station	Writing on a men's toilet wall
Photo_01.jpg	22/02/2019	Matt Miller	Tampere railway station	Writing on station platform wall

Note. Adapted from Example 7 in *Processing qualitative data files*, by Finnish Social Science Data Archive, n.d. c.

While dedicated tools for qualitative data documentation remain less developed than those available for quantitative data, several useful resources exist to support researchers in this task. The UK Data Service provides a data listing template designed for interview-oriented documentation (UK Data Service, n.d. b) as well as examples of completed data lists covering a range of qualitative material types (UK Data Service, n.d. a). Another valuable starting point is the Interview Metadata and Transcript Template developed by the Dutch National Centre of Expertise for Research Data Archiving (DANS) and partners, which provides a structured framework for documenting interviews alongside transcript files (DANS, 2023). Researchers working with qualitative data in dedicated software environments such as ATLAS.ti, NVivo, or MAXQDA will also find that these tools include built-in features for adding descriptive information, annotations, and memos directly to qualitative materials — a form of embedded documentation.

Once data have been documented at the variable or material level, the final step in the documentation journey is to make them visible and accessible to the broader scientific community. This is the purpose of the deposit level.

3.4 THE DEPOSIT LEVEL — THE DIGITAL STOREFRONT FOR THE SCIENTIFIC COMMUNITY

Unlike project- or file-level documentation, which researchers create and maintain according to domain-specific conventions, deposit-level metadata is structured by the repository itself. While the repository provides the framework, researchers supply the content when a research project nears its conclusion. This involves transferring research materials, including data and supporting documentation, from a researcher's working environment to a data repository. At this stage, documentation no longer serves only the research team: it becomes a resource for the broader scientific community. This is referred to as the deposit-level documentation. This layer acts as a digital storefront for research materials, making datasets discoverable, citable and accessible to the global scientific community (Pampel et al., 2013).

Metadata Standards and Controlled Vocabularies

The structure provided by data repositories is built on two complementary pillars: metadata standards and controlled vocabularies (CVs). Metadata standards are formally defined sets of elements, structures, and rules developed by communities or standardisation bodies to ensure that resources are described consistently and in ways that support interoperability across systems and disciplines (Riley, 2017; Zeng & Qin, 2020). In practice, they determine what information researchers must or can provide when depositing data, such as the title, authorship, date of collection, geographic coverage, or methodology. Dublin Core and DataCite are two of the most widely used examples by data repositories. Dublin Core, developed in 1995 and maintained by the Dublin Core Metadata Initiative, proposes fifteen basic descriptive elements — including title, creator, subject, description, and date — that can be applied to any digital resource regardless of discipline or format (Dublin Core Metadata Initiative, 2012). DataCite, developed specifically for research data, offers a richer descriptive framework and supports persistent identifiers such as Digital Object Identifiers (DOIs), which provide stable, citable references to datasets (DataCite Metadata Working Group, 2021). For social science-oriented data repositories, the DDI metadata standard seems to be the most widely used.

Metadata standards alone, however, cannot guarantee consistency across projects. Without a shared vocabulary, “Switzerland” in one dataset might appear as “Swiss”, “Suisse”, “Schweiz”,

or “CH” in another, making it difficult to retrieve or compare studies. This is where Controlled Vocabularies (CVs) come in. CVs are predefined, authorised lists of terms that researchers select when describing their data. When a repository asks to choose keywords from a dropdown menu, you are most likely using a CV. Well-known examples in the social sciences include the CESSDA Topic Classification (CESSDA, 2025), a hierarchical thesaurus of subject terms used by European social science data archives; the UNESCO Thesaurus (UNESCO, n.d.), a multilingual controlled vocabulary covering education, science, culture, and social sciences; and the European Language Social Science Thesaurus (ELSST) (CESSDA and Service Providers, 2025), which provides multilingual subject terms specifically designed for the social sciences and is used by repositories such as SWISSUbase and the UK Data Service.

The DDI: The Standard for Social Sciences

Several metadata standards exist, and the most appropriate one depends on the type of repository and the nature of the data. Generalist standards such as Dublin Core or DataCite are widely adopted for their simplicity and broad applicability. However, for the social sciences, the DDI is considered the gold standard (DDI Alliance, n.d.; Vardigan et al., 2008).

DDI is an international standard designed specifically to describe social, behavioural, and economic research data. Unlike DataCite, which captures only basic bibliographic information (author, title, date), DDI allows researchers to provide a deep, structured description of their research materials. This includes, for example, the full wording of survey questions and response categories, the universe of the population studied, the sampling methodology used, variable-level descriptions and value labels, and the mode of data collection (e.g., online survey, face-to-face interview). By encoding both project-level and data-level documentation within a single, machine-readable framework, DDI bridges the gap between the two. Project-level documentation describes the study as a whole, while data-level documentation describes individual variables and materials (Vardigan et al., 2008).

This level of detail is not merely a formality. It is what allows another researcher, i.e. years later and in a different country, to assess whether a dataset is suitable for their purposes, to understand what a variable labelled "Q12" actually measured, and to replicate or build on the original study. Domain-specific repositories such as SWISSUbase (SWISSUbase, n.d.), the UK Data Service (UK Data Service, n.d. c), and ICPSR (University of Michigan — Institute for Social Research, 2025) use DDI to ensure that social science data remain meaningful and reusable well beyond the life of the original project.

Generalist versus Domain-Specific Repositories: A Strategic Trade-off

Data repositories vary considerably in their scope, governance, and descriptive requirements (Kindling et al., 2017; Pampel et al., 2013). A key distinction is between generalist and domain-specific repositories, each presenting different trade-offs for researchers.

Generalist repositories (e.g., Zenodo, Figshare, Dryad) accept data from any discipline and use lightweight metadata standards such as Dublin Core or DataCite (Pampel et al., 2013). They are quick to use, have a low barrier to entry, and can accommodate a wide range of file types and research outputs. However, because their descriptions are generic, data deposited in these repositories may be harder for specialists to discover. A political scientist searching for survey data on electoral behaviour is unlikely to find it easily in a generalist repository where it is described only by a title and a few keywords.

Domain-specific repositories such as SWISSUbase or UK Data Service are designed for particular disciplines and require more comprehensive metadata (Corti et al., 2014; Kindling et al., 2017). They typically require the use of discipline-specific metadata — such as information on sampling methods, the population universe, or data collection instruments — and often provide professional data curation support (SWISSUbase, n.d.). While the deposit process requires greater effort, the reward is substantially better discoverability among peers and a higher likelihood that the data will be understood and reused correctly.

Therefore, selecting a repository is not a neutral or purely administrative act. Because repositories differ in scope, governance, and descriptive requirements (Kindling et al., 2017; Pampel et al., 2013), the choice of repository is effectively the choice of a descriptive framework — and, by extension, a choice about who will be able to find, understand, and reuse the data. Researchers should thus consider the nature of their data, their target audience, and any funder or institutional requirements before selecting a repository.

4. DOCUMENTATION STRATEGY: KEY RECOMMENDATIONS FOR SOCIAL SCIENTISTS

Taken together, these four levels of documentation — project, file and folder, data, and deposit — form an integrated system rather than a checklist of isolated tasks. Each layer addresses a distinct set of questions and serves a different audience: project-level documentation orients readers to the study's purpose and design; file and folder documentation ensures that research materials are navigable and intelligible; data-level documentation makes individual variables and qualitative records interpretable; and deposit-level documentation connects the work to the broader scientific community.

No single layer is sufficient on its own. A well-described dataset deposited in a domain-specific repository remains difficult to reuse if the files are poorly organised or the variables are unlabelled. Conversely, meticulous internal documentation loses much of its value if the data are never deposited in a findable, citable form. It is the combination of all four that transforms raw data into a durable, trustworthy scientific resource (Borgman, 2017; Wilkinson et al., 2016).

For social scientists in particular, this layered approach is not just good practice; it is essential for meaningful data sharing and responsible research. The complexity of social science data — the interpretive weight of a survey question, the contextual richness of an interview, the sampling decisions that shape a dataset — can only be conveyed through careful, multi-level documentation. Investing in this process is, ultimately, an investment in the long-term scientific value of one's research.

The following recommendations translate these principles into concrete practice. They are not prescriptive rules but adaptable guidelines that researchers can tailor to the specific characteristics of their projects, disciplines, and institutional contexts. Taken together, they reflect a simple underlying conviction: that good documentation is not a bureaucratic burden but an investment in transparency, credibility, and long-term impact of one's research (Borgman, 2017; Corti et al., 2014).

Recommendation 1 — Document with a purpose and for a broad audience

Documentation is an integral part of the research process, not an optional afterthought. It should be created with a clear purpose: to ensure that data can be understood, evaluated, reproduced, and reused by others. Because researchers cannot predict exactly who will need to use their documentation in the future — whether their future self, collaborators, reviewers, or unknown secondary users — it should be designed to support a broad and potentially distant audience (Corti et al., 2014; Whyte & Tedds, 2011).

A practical way to assess whether documentation is sufficient is to apply what might be called the external user test: a dataset is considered well documented if a person not involved in the project can independently understand the research design and context, understand how data were collected, processed, and analysed, know what each variable or data object means, reproduce the data preparation steps, and reuse the data appropriately and responsibly. If a researcher outside the project could not do these things with the documentation available, more documentation is needed (Humphrey, 2014; UK Data Archive, n.d. a).

Recommendation 2 — Prioritise quality, clarity, and proportionality

Good documentation is not about producing as much material as possible. It is about providing information that is accurate, relevant, well-structured, and easy to understand. Both extremes — excessive documentation and insufficient documentation — undermine usability. Overly long or redundant documentation makes data harder to interpret and maintain, increases the risk of inconsistencies, and can overwhelm future users. Conversely, overly minimal documentation fails to provide essential context and undermines reproducibility and reuse (Corti et al., 2014; Humphrey, 2014).

A balanced approach is recommended. Use straightforward language, simple structures, and standard formats, but ensure that all essential methodological, contextual, and structural information is included. Organise documentation so that users can easily locate what they need without navigating unnecessary details and avoid redundant files or repeated information across documents. High-quality documentation is purposeful, coherent, and proportionate to the complexity of the data it describes — detailed where detail is needed, concise where it is not (CESSDA Training Team, 2020; Finnish Social Science Data Archive, n.d. b).

The appropriate level of documentation will also vary depending on the sensitivity of the data, the intended audience, and the requirements of the repository or funder. Data that will be deposited in a domain-specific repository for secondary analysis by other researchers require a particularly thorough documentation.

Recommendation 3 — Begin documenting when you start your project and update it regularly

Documentation is most effective when it is created from the very beginning of a research project and maintained throughout the entire lifecycle of the data. Waiting until the end of the project to document often leads to incomplete, inaccurate, or lost information: methodological decisions, contextual details, and data processing steps become difficult to reconstruct retrospectively, and the tacit knowledge accumulated during fieldwork or analysis fades quickly once a project concludes (Corti et al., 2014; Shankar et al., 2016; Whyte & Tedds, 2011).

Starting early ensures that this tacit knowledge is captured before it disappears, and that documentation evolves naturally alongside the data rather than being assembled under pressure at the point of deposit. In practice, this means creating a README file or project log at the

outset of the project, recording methodological decisions as they are made, updating codebooks or data dictionaries each time variables are added or recoded, and noting any anomalies, corrections, or changes to the data as they occur.

Updating documentation regularly is equally important. As files change, variables are recoded, new data are added, or analytical decisions are revised, documentation must be kept current to remain reliable. Regular updates prevent inconsistencies between the data and their description, support transparent collaboration within the research team, and make it substantially easier to prepare data for publication, archiving, or reuse. Treating documentation as a continuous, iterative practice rather than a one-time task is one of the most effective ways to ensure the long-term integrity and usability of research data (CESSDA Training Team, 2020; Corti et al., 2014).

Recommendation 4 — Use templates, tools, and established standards

Researchers do not need to build their documentation from scratch. A growing range of templates, tools, and standards is available to reduce the burden of documentation, improve consistency and completeness, align descriptions with recognised practices, and make data sharing, long-term preservation, and reuse easier (CESSDA Training Team, 2020; Wilkinson et al., 2016).

At the project level, a user guide or the Cornell University README template (Cornell Data Services, n.d. a) provides a well-structured starting point for describing the overall context and organisation of a research project. At the file and folder level, the same template can be adapted to document directory structures and naming conventions. More comprehensive guidance on organising and documenting research data at every stage of the data lifecycle is also available through the CESSDA Data Management Expert Guide (CESSDA Training Team, 2020), developed specifically for European social science researchers and covering documentation from project planning through to archiving and publication. At the data level, codebook templates are available through software such as SPSS, Stata, and R for quantitative data. For researchers seeking a more structured and interoperable approach to codebook creation, the DDI Codebook standard (DDI Alliance, 2012) provides an XML-based template for documenting variables, file structures, and study-level information in a format that is directly compatible with major social science repositories. For qualitative data, the UK Data Service data listing template (UK Data Archive, n.d. c) and the DANS Interview Metadata and Transcript Template (DANS, 2023) offer practical frameworks for documenting interviews, photographs, and other non-numeric materials. At the repository level, domain-specific repositories such as SWISSUbase, the UK Data Service, and ICPSR provide structured metadata entry forms guided by standards such as DDI, which itself serves as a template for comprehensive social science data description (DDI Alliance, n.d.; Vardigan et al., 2008).

Beyond templates, version control tools such as GitHub, Microsoft SharePoint, and Google Workspace can help researchers manage and track changes to documentation files over time. Data management planning tools such as DMP Online (Digital Curation Centre, n.d.), or the SNSF Data Management Plan template (Swiss National Science Foundation, n.d.) also encourage researchers to think about documentation needs from the very start of a project, before data collection begins.

Recommendation 5 — Adapt documentation to your data and context

No single template or documentation approach can cover all situations. Documentation must be adapted to the specific characteristics of the research project, the nature of the data produced, and the context in which they will be shared or reused. Because data differ in type, structure, discipline, and intended use, the amount, form, and granularity of documentation should reflect the methodological decisions that shaped the data and the expectations of the audiences who will use them (Corti et al., 2014; Humphrey, 2014).

Survey data, qualitative interviews, administrative datasets, experimental measures, and multimodal materials each require different kinds of contextual, technical, and methodological information. A longitudinal panel study will require documentation of how the sample evolved across waves and how variables were harmonised over time. An interview-based qualitative study will require detailed information about context and session. A dataset derived from administrative records will require documentation of the source, any transformations applied, and the conditions under which access was granted.

Documentation requirements also vary according to disciplinary conventions, repository requirements, and funder mandates. Researchers should therefore familiarise themselves with the documentation standards and expectations of their target repository before beginning to prepare their data for deposit and should treat those standards not as bureaucratic constraints but as a shared language that makes their data legible to the broadest possible community of users (Kindling et al., 2017; Pampel et al., 2013).

Conclusion

These five recommendations share a common thread: good documentation is not something that happens to data, but something that is built into the research process from the outset. It requires intention, consistency, and a willingness to think beyond the immediate needs of the research team towards the needs of future users who may encounter the data in entirely different contexts. When documentation is treated as an integral part of research practice — rather than an administrative obligation to be discharged at the end of a project — it becomes a powerful tool for transparency, reuse, and scientific impact. In the social sciences, where data are rarely self-explanatory and where the conditions of their production are inseparable from their meaning, this investment is not optional: it is a condition of responsible and credible research (Borgman, 2017; Corti et al., 2014; Wilkinson et al., 2016).

FURTHER READINGS AND USEFUL WEB LINKS

General Overviews (all Levels)

Chapter 2 ("Organise and Document") and Chapter 3 ("Process") of the CESSDA Data Management Expert Guide, developed by the CESSDA Training Team (2017–2022), provide a good overview of documentation practices across all levels, with useful examples for both quantitative and qualitative data. The full guide is available online at https://dmeg.cessda.eu and as an offline PDF via Zenodo (https://zenodo.org/records/3820473).

The UK Data Archive provides a complete training module on best practices for documenting data collections: https://zenodo.org/records/3820473. Their learning hub also contains dedicated web pages on data-level documentation, metadata, study-level documentation, and

documentation resources: https://ukdataservice.ac.uk/learning-hub/research-data-management/#document-your-data.

Project Level

Cornell Data Services provides the most widely used practical README template for describing the overall context and organisation of a research project: https://data.research.cornell.edu/data-management/sharing/readme/.

File and Folder Level

The UK Data Service provides guidance on file naming conventions and folder organisation as part of its research data management learning hub: https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/organising/.

Data Level

Quantitative data

ICPSR's Guide to Social Science Data Preparation and Archiving (5th ed., 2012) is a comprehensive reference for codebook creation, variable documentation, and data cleaning, available freely as a PDF: https://www.icpsr.umich.edu/files/deposit/dataprep.pdf.

The Finnish Social Science Data Archive (FSD) website provides detailed guidelines on quantitative data documentation, including variable labelling, codebook structure, and file preparation: https://www.fsd.tuni.fi/en/services/data-management-guidelines/processing-quantitative-data-files/.

The DDI Alliance's getting-started page for DDI Codebook offers practical guidance on using the DDI standard to structure codebooks independently of a repository deposit: https://ddialliance.org/getting-started-ddi-c.

Qualitative data

The Finnish Social Science Data Archive guidelines also cover qualitative data documentation in detail, including interview transcripts, field notes, and visual materials: https://www.fsd.tuni.fi/en/services/data-management-guidelines/processing-qualitative-data-files/.

The UK Data Service's dedicated page on data-level documentation addresses qualitative materials specifically: https://ukdataservice.ac.uk/learning-hub/research-data-management/document-your-data/data-level/.

Repository Level

SWISSUbase resources

SWISSUbase is the national Swiss research data platform recommended by the SNSF for social science data deposit. The following resources are available directly from SWISSUbase and FORS:

- SWISSUbase main platform and public catalogue: https://www.swissubase.ch/en/
- FORS Help and Resources page, including webinars on data documentation, anonymisation, and data sharing: https://forscenter.ch/data-services/help-resources/

- SWISSUbase Info website (mission, services, and support): https://info.swissubase.ch
- FORS Guide: *Preparing your social science data for SWISSUbase — in brief* (PDF): https://forscenter.ch/wp-content/uploads/2022/11/the-preparation-of-social-science-data-for-swissubase-in-brief_en.pdf
- FORS Guide: *The preparation of social science data for SWISSUbase — in detail* (PDF): https://forscenter.ch/wp-content/uploads/2022/11/the-preparation-of-social-science-data-for-swissubase-in-detail_en.pdf
- SWISSUbase Metadata Guide (version 2.1, May 2022), describing all metadata fields required at the project and dataset levels (PDF): https://resources.swissubase.ch/wp-content/uploads/2022/05/SWISSUbase-Metadata-Guide_May-2022.pdf

REFERENCES

Adams, D. (1979). *The Hitchhiker's Guide to the Galaxy*. Harmony Books.

Tresch, A., Lauener, L., Bernhard, L., & Scaperrotta, L. (2019). *Swiss Election Study (Selects) 2019*. https://www.swissubase.ch/en/catalogue/studies/13846/16586/overview

Arms, W. Y. (2012). The 1990s: the formative years of digital libraries. *Library Hi Tech*, *30*(4), 579–591.https://doi.org/10.1108/07378831211285068

Banerjee, S. (2025). *Decoding Metadata | openscience.eu*. https://www.openscience.eu/article/infrastructure/decoding-metadata

Bartling, S., & Friesike, S. (2014). Opening science. In S. Bartling & S. Friesike (Eds.), *Opening Science*. Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8

Bellier-Chaussonnier, M. (2002). Representations of libraries in classical Greece. *Revue des études anciennes*, *104*(3–4), 329–347. https://doi.org/10.3406/REA.2002.4873

Berkowitz, H., & Delacour, H. (2022). Opening research data: What does it mean for social sciences? *Management (France)*, *25*(4), 1–15. https://doi.org/10.37725/mgmt.v25.9123

Borgman, C. L. (2017). Big data, little data, no data: Scholarship in the networked world. *Canadian Journal of Communication, 42*(1), 148–150. https://doi.org/10.22230/cjc.2017v42n1a3152

Bowker, G. C. (2001). The Intellectual Foundation of Information Organization. MIT Press, Cambridge, Massachusetts and London, England, 2000, pp. xiv+ 255, ISBN 0-262-19433-3 [Book review]. *Information Processing & Management, 37*(5), 763–764. https://doi.org/10.1016/S0306-4573(01)00022-X

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science*, *42*(5), 351–360. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3

CESSDA and Service Providers (2025). *The European Language Social Science Thesaurus (ELSST)*. Retrieved December 9, 2025, from https://elsst.cessda.eu

CESSDA. (2025). *CESSDA Topic classification*. Version 4.2.3. Retrieved December 9, 2025, from https://vocabularies.cessda.eu/vocabulary/TopicClassification?tab=export

CESSDA Training Team. (2020). *CESSDA data management expert guide*. https://doi.org/10.5281/zenodo.3820473

Cornell Data Services. (n.d. a). *Writing READMEs for research data*. Retrieved December 9, 2025, from https://data.research.cornell.edu/data-management/sharing/readme/

Cornell Data Services. (n.d. b). *File management*. Retrieved December 9, 2025, from https://data.research.cornell.edu/data-management/storing-and-managing/file-management/

Corti, L., Eynden, V. van den, Bishop, L., & Woollard, M. (2014). *Managing and sharing research data: a guide to good practice*. Sage Publishing Book Webpage.

Data Archiving and Networked Services DANS. (2023). *Interview metadata and transcript template*. Retrieved December 9, 2025, from https://www.uva.nl/binaries/content/assets/faculteiten/faculteit-der-geesteswetenschappen/disciplines/conservering-en-restauratie/interviews-in-conservation-initiative/en/uva_dans_template_interview_metadata-transcript.docx

DataCite Metadata Working Group. (2021). *DataCite metadata schema documentation for the publication and citation of research data and other research outputs*. https://doi.org/10.14454/3w3z-sa82

DDI Alliance. (2012). *DDI codebook version 2.5*. Retrieved December 9, 2025, from https://ddialliance.org/ddi-codebook

DDI Alliance. (n.d.). *Data Documentation Initiative*. Retrieved December 9, 2025, from https://ddialliance.org

Digital Curation Centre. (n.d.). *DMP Online*. Retrieved December 9, 2025, from https://dmponline.dcc.ac.uk

Dublin Core Metadata Initiative. (2012). *Dublin core metadata element set, Version 1.1*. Retrieved December 9, 2025, from https://www.dublincore.org/specifications/dublin-core/dces/

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science, 41(5), 667–690*. https://doi.org/10.1177/0306312711413314

Finnish Social Science Data Archive. (n.d.a). *Description of data files*. Retrieved December 9, 2025, from https://www.fsd.tuni.fi/en/services/data-management-guidelines/data-description-and-metadata/#description-of-data-files

Finnish Social Science Data Archive. (n.d.b). *Description of how the study was conducted*. Retrieved December 9, 2025, from https://www.fsd.tuni.fi/en/services/data-

management-guidelines/data-description-and-metadata/#description-of-how-the-study-was-conducted

Finnish Social Science Data Archive. (n.d.c). *Processing qualitative data files*. Retrieved December 9, 2025, from https://www.fsd.tuni.fi/en/services/data-management-guidelines/processing-qualitative-data-files/

Finnish Social Science Data Archive. (n.d.d). *Processing quantitative data files*. Retrieved December 9, 2025, from https://www.fsd.tuni.fi/en/services/data-management-guidelines/processing-quantitative-data-files/

FORS. (2021). *ch-x 2016/2017: Read me (list of files)* [ReadMe file]. SWISSUbase. https://www.swissubase.ch/en/catalogue/studies/13313/latest/datasets/1107/1767/files/17296/24511/physicalFile

FORS. (n.d.). *Template: Interview transcript* [Word document]. https://forscenter.ch/wp-content/uploads/2022/09/modele-transcription-entretien_en.docx

Gilliland, A. J., Gill, T., Woodley, M. S., & Whalen, M. (2008). *Metadata and the web. Introduction to metadata* (2nd ed.). Getty Research Institute, 20–37.

Greenberg, J. (2003). Metadata and the world wide web. In M. A.Drake (Ed.), Encyclopedia of library and information science (2nd ed., pp. 1876–1888). Marcel Dekker.

Guilleminot-Chrétien, G. (1989). Histoire des bibliothèques françaises, tome II: les bibliothèques sous l'ancien régime. 1530-1789. Paris, Promodis-Éditions du Cercle de la Librairie, 1988, XV, 547 pages. *Bulletin du bibliophile*, (1), 186–188.

Hauptmann, E. (2024). Rediscovering the early history of social scientific data centers in the US. *Revue d'histoire des sciences humaines*, (45), 43–67. https://doi.org/10.4000/13AUZ

Heers, M., & Hupka-Brunner, S. (2023). *Parental investment in children's education (PICE) - Technical report*. SWISSUbase. https://www.swissubase.ch/en/catalogue/studies/20043/19785/overview

European Commission. (n.d.). Retrieved March 24, 2026, from https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en

Humphrey, C. (2014). Metadata in the social sciences. *Encyclopedia of Quality of Life and Well-Being Research*, 4002–4004. https://doi.org/10.1007/978-94-007-0753-5_1795

Inter-University Consortium for Political and Social Research ICPSR. (2012). *Guide to social science data preparation and archiving: Best practice throughout the data life cycle* (6th ed.). *https://www.icpsr.umich.edu/sites/icpsr/manage-data/prepare-deposit/guide*

Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., Witt, M., Schirmbacher, P., Bertelmann, R., & Scholze, F. (2017). The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine*, *23*(3–4). https://doi.org/10.1045/MARCH2017-KINDLING

Lerner, F. (2001). *The story of libraries: from the invention of writing to the computer age*. Bloomsbury Publishing. https://archive.org/details/storyoflibraries0000lern

McCallum, S. H. (2002a). International MARC: Past, present, and future. *Advances in Librarianship, 26*, 127–148. https://doi.org/10.1016/S0065-2830(02)80023-X

McCallum, S. H. (2002b). MARC: Keystone for library automation. *IEEE Annals of the History of Computing*, 24(2), 34–49. https://doi.org/10.1109/MAHC.2002.1010068

Mitra, P. (2016). Dewey decimal system. *Encyclopedia of Database Systems*, 1–2. https://doi.org/10.1007/978-1-4899-7993-3_877-2

Niu, J. (2009). *Perceived documentation quality of social science data*. University of Michigan. Retrieved March 24, 2026, from https://hdl.handle.net/2027.42/63871

Niu, J., & Hedstrom, M. (2008). Documentation evaluation model for social science data. *Proceedings of the ASIST Annual Meeting*, *45*. https://doi.org/10.1002/MEET.2008.1450450223

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P., & Dierolf, U. (2013). Making research data repositories visible: the re3data.org registry. *PloS One*, *8*(11), e78080. https://doi.org/10.1371/journal.pone.0078080

Rasmussen, K. B. (2014). Social science metadata and the foundations of the DDI. *IASSIST Quarterly*, *37*(1–4), 28–28. https://doi.org/10.29173/IQ499

Riley, J. (2017). *Understanding metadata*. National Information Standards Organization NISO, 23, 7-10. https://www.niso.org/publications/understanding-metadata-2017

Selects. (2021). *Post-Election Survey Codebook - EN* [Codebook]. SWISSUbase. https://www.swissubase.ch/en/catalogue/studies/13846/latest/datasets/1179/1875/files/18856/25847/physicalFile

Shankar, K., Eschenfelder, K. R., & Downey, G. (2016). Studying the history of social science data archives as knowledge infrastructure. *Science & Technology Studies, 29*(2), 62–73. https://doi.org/10.23987/sts.55691

Sherman Center for Digital Scholarship. (n.d.). *Organization and documentation*. Retrieved September 2, 2026 from https://learn.scds.ca/rdm-best-practices/topics/2-organization-and-documentation.html

Swiss National Science Foundation. (n.d.). (2026, August 24). *SNSF Open Research Data: Data management plan*. SNSF. https://www.snf.ch/en/FAiWVH4WvpKvohw9/topic/open-research-data

SWISSUbase. (n.d.). *SWISSUbase Repository*. Retrieved March 24, 2026, from https://www.swissubase.ch/en/

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS One, 6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

TREE. (2024). *Working with the TREE2 data release: How to get started*. SWISSUbase. https://www.swissubase.ch/fr/catalogue/studies/12476/latest/datasets/1255/3340/files/18970/27630/physicalFile

UK Data Archive. (n.d. a). *Study-level documentation*. Retrieved March 24, 2026, from https://ukdataservice.ac.uk/learning-hub/research-data-management/document-your-data/study-level-documentation/

UK Data Archive. (n.d. b). *UK — Data-level documentation*. Retrieved March 24, 2026, from https://ukdataservice.ac.uk/learning-hub/research-data-management/document-your-data/data-level/

UK Data Archive. (n.d. c). *UK — Qualitative data documentation*. Retrieved March 24, 2026, from https://ukdataservice.ac.uk/learning-hub/research-data-management/document-your-data/data-level/data-documentation-qualitative-data/

UK Data Archive. (n.d. d). *UK Data documentation: quantitative data*. Retrieved March 24, 2026, from https://ukdataservice.ac.uk/learning-hub/research-data-management/document-your-data/data-level/data-documentation-quantitative-data/

UK Data Service. (n.d. a). *Completed data list examples. UK Data Service*. Retrieved March 24, 2026, from https://ukdataservice.ac.uk/app/uploads/datalist-1.gif

UK Data Service. (n.d. b). *Data listing template. UK Data Service*. Retrieved March 24, 2026, from https://ukdataservice.ac.uk/app/uploads/uk_data_archive_data_listing_template.xlsx

UK Data Service. (n.d. c). *Deposit Data*. Retrieved March 24, 2026, from https://ukdataservice.ac.uk/deposit-data/

UNESCO. (n.d.). *UNESCO Thesaurus*. Retrieved March 24, 2026, from https://skos.um.es/unesco6

University of Michigan — Institute for Social Research. (2025). *ICPSR Repository*. Retrieved March 24, 2026, from https://www.icpsr.umich.edu/sites/icpsr/home

Van den Eynden, V., & Chatsiou, K. (2011). *Data management for qualitative data using NVIVO 9*. https://ukdataservice.ac.uk//app/uploads/ukda-datamanagement-nvivo.pdf

van Zuilen, M., Ruiz, J. G., & Mintzer, M. (2006). Guidelines for developing user guides and facilitators manuals for educational resources. *MedEdPORTAL*. https://doi.org/10.15766/mep_2374-8265.169

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, *3*(1), 107–113. https://doi.org/10.2218/IJDC.V3I1.45

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. https://doi.org/10.1016/J.JBUSRES.2017.12.043

Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). *Dublin core metadata for resource discovery*. https://doi.org/10.17487/RFC2413

Whyte, A., & Tedds, J. (2011). *Making the case for research data management*. Digital Curation Centre. Retrieved March 24, 2026, from https://www.dcc.ac.uk/guidance/briefing-papers/making-case-rdm

Wilkinson, M., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

Woodley, M. S. (2008). *Crosswalks, metadata harvesting, federated searching, metasearching: Using metadata to connect users and information.* https://scholarworks.calstate.edu/downloads/5999n6527

Yang, W., Fu, R., Amin, M. B., & Kang, B. (2025). The impact of modern AI in metadata management. *Human-Centric Intelligent Systems*, *5*(3), 323–350. https://doi.org/10.1007/s44230-025-00106-5

Zeng, M. L., & Qin, J. (2020). *Metadata*. American Library Association. ISBN: 1555706355, 9781555706357