TU WIEN Informatics

Digital Amnesia and Collective Memory

A Platform-Centred YouTube Analysis — A Digital Humanism Mandate

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Amer Alkojjeh, BSc.
Registration Number 01646631

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. A Min Tjoa

Vienna, July 1, 2026

Amer Alkojjeh	A Min Tjoa

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Declaration of Authorship

Amer Alkojjeh, BSc.

I hereby declare that I have written this Thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work - including tables, maps and figures - which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.
I further declare that I have used generative AI tools only as an aid, and that my own intellectual and creative efforts predominate in this work. In the appendix "Overview of Generative AI Tools Used" I have listed all generative AI tools that were used in the creation of this work, and indicated where in the work they were used. If whole passages of text were used without substantial changes, I have indicated the input (prompts) I formulated and the IT application used with its product name and version number/date.

Vienna, July 1, 2026

Amer Alkojjeh

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acknowledgements

This thesis was completed at TU Wien and would not have been possible without the support of a number of people and organisations.

I thank my supervisor for the sustained guidance, critical engagement, and patience throughout a research process that was rarely straightforward. The feedback offered at each stage — on framing, method, and argument — shaped this work in ways that a brief acknowledgement cannot adequately capture.

I thank those who took the time to discuss the ideas developed here, whether in formal settings or in passing conversation. Research on platform governance, digital memory, and web archiving sits at the intersection of several fields, and these discussions helped clarify the boundaries and implications of the work.

I thank the Internet Archive for maintaining the publicly accessible Wayback Machine and its archival services. Without these public archival resources, the empirical basis of this thesis could not have been constructed.

I thank SyrianMemory, the non-governmental organisation that provided the dataset used in this thesis. By collecting and classifying Syria-related YouTube records, SyrianMemory provided the empirical foundation without which this research could not have been carried out.

I thank the people who archived videos of the Syrian conflict, often under risk and without institutional support, for creating a record that would otherwise not exist. This work takes their efforts seriously, and any contribution it makes to the preservation and accountability infrastructure that surrounds such records belongs, in part, to them.

Finally, I thank Nadja Thoma for her support throughout this process, including during long periods of uncertainty, pressure, and delay.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Abstract

Digital platforms function as central infrastructures for the production and circulation of digital records that shape contemporary collective memory. At the same time, the long-term availability of platform-hosted content is unstable and difficult to assess from an external perspective due to limited transparency and heterogeneous archival traces. This thesis examines these conditions through an empirical study of Syria-related YouTube videos.

Using externally observable platform-facing checks and web archive evidence, the study analyses a corpus of over 50,000 videos collected between 2011 and 2025. A rule-based observational measurement system is implemented to reconstruct per-video availability trajectories and to derive analytical feature matrices for descriptive analysis, time-to-event descriptive analysis, and regression modelling.

The results show that most videos with archival history remain empirically indeterminate within the observation window, with explicitly observed unavailability limited to a small subset of cases. Statistical analysis identifies conditional associations between content attributes and final platform-observed availability outcomes, but these associations remain fragmented and constrained by limits of observability. To address these conditions, the thesis introduces a system-level artefact that represents graded states of archival observability and preserves incomplete and partial traces as explicit analytical outcomes.

By documenting the limits of external observation across platform and archival layers and by operationalising these limits within a reproducible measurement framework, the thesis contributes empirical and methodological foundations for the study of platform-mediated digital memory under conditions of platform and archival opacity.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Abstract **vii**

Contents **ix**

1 Introduction **1**
1.1 Background and Motivation 1
1.2 Problem Statement 2
1.3 Research Gap 2
1.4 Research Questions 4
1.5 Objectives and Contributions 5
1.6 Structure of the Thesis 6

2 Theoretical Background and Related Work **7**
2.1 Introduction 7
2.2 Collective Memory and Digital Forgetting 8
2.3 Syrian Conflict as Platform-Mediated Memory 9
2.4 Conditions of Remembrance 12
2.5 Processes of Forgetting 13
2.6 Digital Humanism as Research Foundation 14

3 Methodology **21**
3.1 Introduction 21
3.2 Research Framework: Design Science Research (DSR) 22
3.3 Relevance Cycle: Problem Grounding 23
3.4 Rigor Cycle: Knowledge Base 28
3.5 Design Cycle: Artefact Design 33
3.6 Design Cycle: Artefact Evaluation 66
3.7 Reproducibility and Ethical Considerations 69
3.8 Contributions and Limitations 70

4 Empirical Analysis **73**
4.1 Analytical Scope and Execution Context 73
4.2 Corpus Definition, Units, and Data State 73
4.3 Wayback CDX Temporal Evidence 74

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.4 Wayback HTML Snapshot Evidence . . . 77
4.5 YouTube Platform Availability States . . . 81
4.6 Availability Histories from Archived Evidence . . . 83
4.7 Cross-Source Observability Boundaries . . . 88
4.8 Content Taxonomy Coverage and Feature Availability . . . 92
4.9 Feature Freeze and Statistical Input Dataset . . . 94
4.10 Summary and Analytical Transition . . . 95

5 Statistical Analysis and System Evaluation 97
5.1 Purpose and Analytical Scope . . . 97
5.2 Analytical Datasets and Feature Matrices . . . 97
5.3 RQ1: Availability Patterns . . . 100
5.4 RQ2 Regression Analysis of Availability Outcomes . . . 105
5.5 RQ3: Design Response . . . 107
5.6 Evaluation of the Observational Measurement System . . . 110
5.7 Summary of Findings and Evaluation . . . 115

6 Discussion 117
6.1 Purpose and Scope of the Discussion . . . 117
6.2 Empirical Findings on Observability Limits . . . 117
6.3 Relationship to Existing Conceptual Frameworks . . . 119
6.4 Observability Conditions Across Frameworks . . . 122
6.5 Implications for Digital Memory Research . . . 122
6.6 Limitations . . . 124
6.7 Summary . . . 124

7 Conclusion 125
7.1 Summary of Findings . . . 125
7.2 Contributions . . . 125
7.3 Future Research . . . 126

A Overview of Generative AI Tools Used 131

B Methodological Supplement 133

C Future Research Project Concepts 137

D Archive-Side Evidence-State Model 141

E Completeness Artefact 155

F Taxonomy Mapping Tables 167

List of Figures 171

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

List of Tables **173**

List of Algorithms **177**

Glossary **179**

Acronyms **189**

Bibliography **191**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

1.1 Background and Motivation

Digital platforms have become central infrastructures for the production and circulation of digital records that shape contemporary collective memory [Gil18, Hos17]. In particular, video-sharing services host large volumes of user-generated material documenting political events, social movements, and armed conflicts. In contexts of political violence, such platforms have functioned as primary channels through which audiovisual traces are published and accessed.

The empirical domain examined in this thesis is the Syrian conflict, which began in 2011 with a popular uprising and subsequently developed into a protracted armed conflict whose effects extended across multiple actors, territories, and institutions [DH18]. The Syrian conflict occupies a distinctive position in the history of platform-mediated documentation: it has been described as "the first YouTube conflict in the same way that Vietnam was the first television conflict" [Whi24], reflecting the degree to which user-generated video uploaded to social media platforms became the primary channel through which events were documented and circulated, particularly in contexts where conventional journalistic access was severely constrained. By some estimates, there came to exist more hours of video content from Syria uploaded to YouTube than there had been hours of the conflict itself [Ris19, Tar24]. The corpus underlying this thesis comprises over 800,000 video identifiers assembled by a non-governmental documentation organisation, of which approximately 50,000 are analysed, spanning the full period from 2011 through 2025. This corpus is not merely a large dataset in the methodological sense. It represents a substantial fragment of the audiovisual record through which the Syrian conflict may be remembered, studied, debated, and contested by future communities — a fragment whose continued accessibility cannot be assumed and whose patterns of availability and loss this thesis undertakes to observe and measure. The Vienna Manifesto on Digital Humanism identifies the safeguarding of information dissemination as a central concern

of platform governance, noting that platforms need to be altered to better safeguard the dissemination of information [WLA+19]. This concern takes on concrete historical weight in a context where the documentation of one of the most widely observed armed conflicts of the early twenty-first century is subject to ongoing platform-mediated loss at scale — loss whose extent and character cannot be determined without external observational tools of the kind developed in this thesis.

At the same time, platform-mediated memory is conditioned by instability in availability and accessibility over time [Hos17]. Content may become unavailable due to moderation actions, policy changes, technical failures, or shifts in platform governance [Gil18]. This thesis examines this tension through a Syria-related YouTube corpus and an observational design that combines platform-facing checks with web archive evidence.

1.2 Problem Statement

Despite their growing role as de facto memory infrastructures, major digital platforms offer limited transparency regarding the long-term availability of user-generated content. Decisions affecting visibility, accessibility, or removal are typically opaque, and the reasons for content unavailability are rarely observable to external observers. For researchers and archivists, this opacity complicates any systematic assessment of how much digital memory is preserved, lost, or transformed over time.

Most research on platform governance and content moderation has focused on legal and policy questions, whereas fewer empirical studies provide systematic methods for observing how long content remains available over time [Gil18, KL20, Zub18, STvO+18]. As a result, there is a lack of reproducible, system-oriented methods that allow researchers to quantify archival persistence and loss without relying on insider access to platform decision-making processes.

This problem is not only technical. When conflict-related digital records become unavailable, inaccessible, or only partly reconstructable, the accessible record of past events becomes smaller. In the Syria-related YouTube case, this affects whether historically relevant traces remain available for collective reference, reuse, and long-term preservation. The research problem addressed in this thesis therefore concerns both the observation of content availability and the conditions under which platform-mediated infrastructures shape collective memory.

1.3 Research Gap

Existing scholarship shows that major digital platforms function as key infrastructures through which digital memory is produced, circulated, and preserved. At the same time, these infrastructures operate with limited transparency and restricted external observability [Gil18, Hos17]. While these studies establish the importance of platforms

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

for contemporary memory practices, several methodological problems remain unresolved when researchers attempt to observe and analyse platform-mediated memory empirically.

G1: Limited external observability of moderation decisions. *No continuous, externally accessible method currently exists for observing how content availability changes on platforms over time as a result of moderation decisions.* Research on platform governance and content moderation has repeatedly emphasised the limited transparency of moderation systems and the difficulty of obtaining reliable information about content-removal decisions [SWQY19, KL20, GBK20]. Early research in this area noted that meaningful transparency requires detailed information about the moderated content, the rule that was applied, and the actor or automated system responsible for the decision [SWQY19].

Recent regulatory developments in the European Union have partially addressed this issue. Under the Digital Services Act, platforms must issue Statements of Reasons for certain moderation decisions, and these statements are collected in the DSA Transparency Database [TFC25, Eur]. However, such regulatory reporting provides information about individual moderation actions rather than continuous observations of content availability. As a result, external empirical research still faces limitations when attempting to observe how content appears, disappears, or changes accessibility on platforms over time.

The invisibility of these decision processes to external observers is not incidental but structural: as the Vienna Manifesto observes, many of the choices embedded in algorithmic systems "remain hidden in software programs implementing algorithms that remain invisible" [WLA+19]. The absence of a continuous external observation method is therefore not merely a gap in the research literature but a consequence of the structural opacity that the Manifesto identifies as a defining feature of contemporary digital infrastructure.

G2: Incomplete and heterogeneous external evidence. *No established methodological approach exists for combining heterogeneous external evidence sources into a coherent and comparable analytical framework for studying platform content.* Researchers who study platforms often rely on external evidence sources such as web archives, platform APIs, or scraped datasets. However, these sources provide incomplete representations of platform content. Web archives preserve snapshots of the web but do not capture all pages or all time periods [HKS+15]. Furthermore, archived web material is not a neutral record of past web content but a constructed artefact shaped by selection, preservation practices, and technical constraints [Bru18].

Platform interfaces introduce additional constraints. Data retrieved through APIs reflects the structure and limits imposed by the platform itself [AK20]. As a result, external datasets often differ in structure, scope, and completeness. While many studies rely on individual sources of platform data, combining heterogeneous sources remains methodologically difficult because different datasets capture different aspects of platform activity and are not directly comparable [CD21, Vic23, RSM22].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

G3: Non-binary trace states and unstable persistence. *Although theoretical and regulatory frameworks recognise multiple intermediate states of content availability, empirical research has not yet operationalised these insights into systematic measurement frameworks for platform-mediated digital traces.* Digital memory research further emphasises that digital traces do not exist in a stable or binary state. Instead, digital content may decay, disappear, or become temporarily inaccessible over time [Hos17]. In this perspective, remembering and forgetting are not fixed states but ongoing processes that involve varying degrees of persistence and accessibility.

Empirical studies of online information persistence have therefore proposed multiple indicators for describing the stability of digital traces, including measures of completeness, representativity, similarity, change over time, and retrievability [Zub18, STvO+18]. At the same time, contemporary platform governance frameworks recognise multiple forms of content restriction. These include removal, disabling access, demotion, labelling, age restriction, and account-level sanctions rather than a single binary distinction between available and removed content [Eur22, Eur]. These observations suggest that platform-mediated digital traces may pass through several observable states rather than a single stable condition.

G4: Reproducibility and methodological transparency in observational research. *No reproducible observational framework currently exists that embeds transparency and accountability principles into the systematic study of how platform content changes availability over time using heterogeneous external traces.* Digital Humanism highlights the importance of transparency and accountability in digital systems and in the research practices used to study them [WPLG22]. Several recent studies in platform and digital-memory research already implement transparent research practices such as data-availability statements, open datasets, and publicly accessible code [BDMN24, BM23, Adr24].

However, these studies primarily investigate discourse, conversational remembering, or mnemonic communities rather than the long-term behaviour of platform content itself. As a result, they do not provide methodological frameworks for systematically observing how platform content changes availability over time using heterogeneous external traces.

Taken together, these limitations indicate that the current literature does not yet provide a reproducible observational framework that can combine heterogeneous external evidence to analyse the persistence and accessibility of platform-mediated digital traces over time while explicitly accounting for incomplete observations and measurement uncertainty.

1.4 Research Questions

Guided by these considerations, the thesis investigates the following research questions:

- **RQ1:** What externally observable patterns of deletion affect Syria-related videos on YouTube?

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- **RQ2:** How do factors such as topic, language, and metadata correlate with the likelihood of platform content removal on YouTube?
- **RQ3:** Under external observability limits, which system-level design improvements can support digital memory preservation and alignment with Digital Humanism?

1.5 Objectives and Contributions

The primary objective of this thesis is to develop and empirically apply a transparent and reproducible observational approach for studying platform-mediated digital memory persistence using externally observable evidence. The study is explicitly designed to operate under conditions of platform opacity and without access to internal platform decision-making processes. To achieve this objective, the thesis makes three interrelated contributions.

Empirical contribution. The thesis provides an empirical analysis of item-level availability trajectories for Syria-related YouTube videos over an extended observation window. Existing research has not produced longitudinal, item-level records of content availability constructed from external evidence sources alone, leaving the question of how specific content items persist, disappear, or reappear over time empirically unaddressed (**G1**). By combining live platform-facing availability checks with web archive evidence, this contribution demonstrates that such trajectories can be reconstructed without access to internal platform decision-making processes, and produces a longitudinal dataset of item-level availability states as a concrete empirical output.

Methodological contribution. The thesis introduces a methodological framework that addresses two related but distinct problems left unresolved in the existing literature. The first is the absence of a principled approach for combining heterogeneous external evidence sources - web archive indices, replayed snapshots, and live platform states - into a coherent analytical dataset, given that these sources differ in structure, scope, and completeness and are not directly comparable (**G2**). The second is the predominance of binary availability classifications in empirical research, which fails to capture the intermediate and unstable states through which platform-mediated digital traces actually pass (**G3**). The framework addresses both problems by operationalising content availability as a graded observational construct, explicitly accounting for partial traces, right-censoring, and cross-source non-alignment, and enabling systematic analysis of how content attributes such as topic, language, and metadata associate with observed availability outcomes.

Practical contribution. The thesis contributes a system-oriented research artefact in the form of a reproducible observational measurement system that embeds transparency, provenance recording, and auditability into the research infrastructure itself (**G4**). Unlike existing studies that implement open research practices as supplementary measures, this

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

artefact embeds transparency into the measurement process itself: recording evidence provenance explicitly, producing auditable analytical datasets whose derivation conditions are independently verifiable, and preserving graded visibility states rather than imputing completeness. In doing so, it makes the evidentiary conditions of platform and archival infrastructure observable and analytically tractable, supporting the identification of longitudinal patterns of preservation and loss (**RQ1**), comparative analysis of differential availability (**RQ2**), and the specification of system-level design considerations grounded in observable infrastructural limitations (**RQ3**).

Together, these contributions support rigorous empirical research on platform-mediated digital memory while making the conditions and limits of observation explicit, reproducible, and analytically tractable. The alignment between the identified research gaps and these contributions is summarised in Table 1.1.

Table 1.1: Alignment between research gaps (G1–G4) and thesis contributions

Research Gap	Empirical	Methodological	Practical
G1	Direct	Direct	Indirect
G2	–	Direct	Indirect
G3	–	Direct	–
G4	–	Indirect	Direct

Direct	primary coverage of a research gap
Indirect	infrastructural support or enabling design features
–	not addressed

1.6 Structure of the Thesis

Chapter 2 introduces the theoretical background, situates digital forgetting within debates on collective memory and platform governance, and defines the role of Digital Humanism as the research motivational foundation of the thesis. Chapter 3 details the research methodology and describes the system design, data sources, and operational measurement system. Chapter 4 reports the empirical observations produced by executing the measurement process and characterises the observable properties and limits of the resulting data. Chapter 5 reports the quantitative analyses for **RQ1** and **RQ2**, formulates the system-level design response for **RQ3**, and evaluates the observational measurement system across all three research questions. Chapter 6 consolidates the findings from Chapters 4 and 5, clarifies their implications under conditions of platform and archival opacity, and documents limitations. Chapter 7 concludes the thesis and outlines directions for future research. Technical documentation and supplementary material are provided in the appendices.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

Theoretical Background and Related Work

2.1 Introduction

Digital platforms increasingly function as infrastructures of collective memory, mediating how contemporary events are recorded, circulated, and preserved over time. In contexts of political violence and armed conflict, these platforms play a particularly important role, as they host large volumes of audiovisual material produced by eyewitnesses, journalists, and civil-society actors. Such material often functions as the major form of digital documentation for events that are otherwise documented only to a limited extent or whose accounts are contested [Ass11].

At the same time, the persistence of digital content on large-scale platforms cannot be taken for granted. Platform governance regimes, automated moderation systems, and evolving policy frameworks introduce instabilities that influence the long-term availability of online material [Gil18]. As a result, digital memory within platform environments is mediated by platform-related structures and exhibits inherent fragility [Gil18]. Existing research has highlighted that processes of disappearance, suppression, or degradation are rarely transparent and often unfold gradually rather than as clearly identifiable events [Gil18].

Theoretical work on collective memory and digital archives has extensively discussed these dynamics, emphasising that remembrance is contingent upon a range of technical, institutional, and socio-cultural conditions [WH25]. Likewise, scholarship on digital forgetting has stressed that loss and erasure in online environments are frequently non-binary, processual, and unevenly distributed [WH25]. However, while these bodies of literature offer valuable conceptual insights, they also point to a central limitation: many

of the mechanisms that shape digital remembrance and forgetting remain empirically opaque, particularly in the context of proprietary platform infrastructures [WH25].

This chapter develops the theoretical foundation for addressing this limitation. Drawing on research in collective memory studies and platform governance, this chapter delineates the conceptual problem space within which digital remembrance and forgetting can be studied under conditions of platform opacity. Digital Humanism is introduced in this chapter as the research motivational foundation through which the societal relevance of platform-mediated digital infrastructures is framed, while questions of empirical observability and methodological design are addressed in Chapter 3.

2.2 Collective Memory and Digital Forgetting

The concept of collective memory refers to the socially mediated processes through which societies construct, stabilise, and transmit interpretations of the past [Ass11]. Classical memory studies have emphasised that remembrance is not merely an aggregation of individual recollections, but a collective phenomenon shaped by institutions, media, and cultural practices. In digital environments, these processes are increasingly mediated by online platforms that store, rank, and curate large volumes of user-generated content [Ass11].

Digital media infrastructures alter the conditions under which collective memory is formed. Unlike traditional archives, platform-based memory systems are characterised by continuous modification, algorithmic ordering, and dependency on proprietary governance regimes [WH25]. As a result, the persistence of digital artefacts is contingent not only on their initial production, but also on ongoing technical maintenance, policy compliance, and visibility within platform ecosystems. Research in digital memory studies describes digital remembrance as unstable and subject to structural forms of fragility [WH25].

Forgetting, in this context, cannot be understood as a singular act of erasure [WH25]. Rather, scholarship has conceptualised digital forgetting as a processual phenomenon that unfolds over time and across multiple layers of mediation [WH25]. Content may lose visibility without being removed, remain indexed while becoming inaccessible, or persist only in fragmented archival traces. Such non-binary forms of loss challenge traditional distinctions between presence and absence and complicate empirical assessment [WH25].

These dynamics are unevenly distributed. Digital forgetting is shaped by dependencies on digital infrastructures and services - specifically by dependencies on apps, platforms, and networks - and the resulting forgetting ecologies are not universally applicable [WH25]. In settings such as war and commemoration, these dynamics can push conflicts to the periphery of the public attention and contribute to weakening the visibility of past events [WH25].

Taken together, the literature on collective memory and digital forgetting establishes two central insights that are relevant for this thesis. First, digital remembrance is contingent upon socio-technical conditions that extend beyond individual acts of documentation.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Second, forgetting in platform environments is gradual, fragmented, and structurally mediated rather than binary or instantaneous. These insights define the conceptual problem space within which subsequent discussions of remembrance conditions and forgetting processes are situated.

When digital records that document past events become unavailable, partially accessible, or inconsistently archived across platforms and web archives, the ability of groups to access, reference, and reuse those records changes. Over time, repeated instances of such availability loss alter what remains practically retrievable within the digital record of an event. In this thesis, this condition is described as *digital amnesia*. The term refers to a collective-level reduction in the accessible digital record of past events resulting from cumulative digital forgetting within platform and archival systems.

2.3 The Syrian Conflict as a Case of Platform-Mediated Collective Memory at Risk

The theoretical account of digital forgetting and digital amnesia developed in the preceding section acquires concrete historical weight when situated against the empirical domain examined in this thesis. The Syrian conflict, which began in 2011 with a popular uprising against the Assad government and subsequently escalated into a protracted armed conflict involving multiple state and non-state actors, constitutes one of the most extensively documented events in the history of platform-mediated communication [Ris19]. It also constitutes one of the clearest cases in which the conditions for digital amnesia identified in the literature — platform-mediated loss, archival fragmentation, and the gradual erosion of accessible documentation — have been empirically observed and reported by civil society organisations working to preserve the digital record [DH18, Hub22].

This section develops the Syrian conflict as a concrete case through which the theoretical problem of platform-mediated digital amnesia becomes historically and empirically visible. It does so in five steps. First, it explains why the Syrian conflict constitutes a distinctive case of platform-based documentation. Second, it shows why this documentation has collective-memory significance. Third, it outlines the role of civil society preservation initiatives and the evidentiary importance of the material. Fourth, it identifies platform moderation as a direct threat to the continuity of that documentation. Fifth, it shows why this documentation crisis remains ongoing rather than historically closed.

Syria as a platform-documented conflict. The significance of this case for collective memory research does not derive primarily from the scale of the conflict, though that scale is substantial. It derives from the specific conditions under which documentation was produced and preserved. Conventional journalistic access to Syria was severely constrained throughout much of the conflict period, with foreign correspondents facing systematic obstruction and physical danger [Whi24]. As a result, a disproportionate share of the audiovisual record of the conflict was produced not by institutional media but by eyewitnesses, activists, and civil society media offices operating within Syria, who

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

uploaded video material to social media platforms — and to YouTube in particular — as documentation became available [Ris19, DH18]. One consequence of this pattern was noted by practitioners at the time: there came to exist more hours of user-generated video content from Syria than there had been hours of the conflict itself [Ris19, Whi24, Tar24]. The Syrian conflict thus became what one practitioner described as "the first YouTube conflict in the same way that Vietnam was the first television conflict" [Whi24]. For the purposes of collective memory formation, this means that platform-hosted documentation does not merely supplement other archival records of the conflict: for large portions of the event, particularly in areas inaccessible to institutional actors, it constitutes the primary available record.

Collective-memory significance of platform-hosted documentation. This observation has direct implications for the collective memory framework developed in this thesis. As Assmann argues, cultural memory depends upon institutionalised mnemotechnics — material carriers, rituals, and institutions that enable retention, retrieval, and transmission of the past over time [Ass11]. In the Syrian case, the institutional carriers of conventional archival memory were largely absent or inaccessible during the conflict period. Platform-hosted video documentation assumed a structural role that, in other historical contexts, would have been performed by institutional archives, state records, or established journalistic repositories. The implications of this substitution are significant: if platform-hosted documentation is lost, degraded, or rendered inaccessible, there is no institutional alternative that can fully fill the resulting gap. The conditions for digital amnesia — understood as a collective-level reduction in the accessible digital record resulting from cumulative digital forgetting — are therefore not merely theoretical in this context but empirically present and documented.

Preservation practice and evidentiary relevance. Civil society organisations working at the intersection of digital preservation and conflict documentation have been acutely aware of this condition. Organisations including the Syrian Archive — subsequently integrated into Mnemonic — and platforms such as SyrianMemory have undertaken systematic efforts to identify, collect, verify, and preserve conflict-related audiovisual material from social media platforms, precisely because they recognised that platform-hosted documentation was subject to ongoing loss [Syr25, DH18]. The Syrian Archive's methodology explicitly addresses the relationship between documentation and future accountability: the organisation established chain-of-custody protocols for archived footage and developed metadata schemas designed to meet the evidential standards required by international investigative bodies, including the International, Impartial and Independent Mechanism for Syria established by the United Nations General Assembly [Ris19, Whi24]. The evidentiary significance of this material has been recognised at the level of international criminal law, where user-generated content from social media has been cited in reports by the United Nations Independent International Commission of Inquiry on Syria and has informed proceedings before the International Criminal

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Court [Ris19]. The relationship between platform-hosted documentation and transitional justice is therefore not merely conceptual but institutionally established [RT23, Whi24].

Platform moderation as a documentation threat. The threat to this documentation emerged not only from the physical conditions of the conflict but from the operational logic of the platforms on which the documentation was hosted. The single most consequential documented event in this regard occurred in 2017, when YouTube deployed updated automated content moderation systems designed to identify and remove material categorised as extremist or terrorist content. The consequence of this deployment was that more than ten per cent of the videos from Syria posted online since 2011 were removed almost overnight — approximately 400,000 videos, the overwhelming majority of which had been uploaded by activists and civil society actors documenting the conflict rather than by actors promoting violence [Tar24, AJAKD$^+$19, Hub22]. The mechanism of removal was a hashing system operating through the Global Internet Forum to Counter Terrorism, whose opacity rendered it impossible to review or challenge [Hub22, AJAKD$^+$19]. Automated flagging systems were designed to identify content categorised by the platform as potentially terrorist — a category whose operational definition prioritised commercial considerations over the distinction between extremist propaganda and conflict documentation [Tar24]. Content moderation decisions of this kind are made at machine speed, without individual human review of the material affected, and without prior notification to the uploaders or to organisations working to preserve the material [Hub22, AJAKD$^+$19].

This large-scale removal event is also relevant from a Digital Humanism perspective, because it exemplifies the structural opacity and societal consequences of platform-mediated documentation loss that are formalised later in Section 2.6. In this chapter, the Syrian case serves primarily to establish the historical and empirical relevance of platform-mediated memory loss; the detailed Digital Humanism framing is defined in the dedicated section below.

The continuing documentation crisis. The documentation crisis in the Syrian case is not a resolved historical condition but an ongoing process. The corpus analysed in this thesis spans 2011 through 2025, a period during which the conflict itself, its political aftermath, and the efforts of civil society organisations to preserve its documentation all continued to unfold [Syr25]. The availability patterns documented in this thesis — including the preponderance of right-censored trajectories, the high proportion of videos without archival history, and the semantic sparsity of archived HTML representations — are not artefacts of a completed historical event but characteristics of an active documentation environment in which content continues to be produced, moderated, and archived under conditions of ongoing instability. The need for a reproducible, externally observable measurement framework is therefore not merely a methodological preference but a substantive response to a real and continuing documentation crisis.

Taken together, the Syrian case functions in this chapter not merely as an empirical example, but as the concrete setting in which the theoretical problem of platform-mediated

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

digital amnesia becomes historically, institutionally, and methodologically visible.

2.4 Conditions of Remembrance

Remembrance in digital environments is not an automatic consequence of content creation or publication. Rather, it is contingent upon a set of conditions that enable digital artefacts to persist, remain accessible, and retain interpretive relevance over time. Memory studies have long emphasised that remembrance depends on social interaction, institutionalised mnemotechnics, and material carriers, rituals, and institutions [Ass11]. In digital contexts, such supports are mediated through platform governance and the algorithmic organisation of visibility [Gil18]. For analytical clarity, this thesis distinguishes technical, institutional, socio-cultural, and temporal conditions of digital remembrance.

Technical conditions. Technical conditions constitute a foundational layer of digital remembrance. Content must remain stored and retrievable within evolving technological environments [Ass11]. Failures at this level—such as data loss, format obsolescence, or infrastructural change—can undermine persistence independently of deliberate removal. In platform-mediated systems, these technical conditions are linked to maintenance decisions and system updates controlled by platform operators rather than individual users.

Institutional conditions. Institutional conditions further shape the possibilities of remembrance. Platform governance regimes define acceptable forms of content, regulate visibility, and enforce compliance through moderation practices [Gil18]. These regimes are dynamic and often opaque, meaning that the continued presence of digital artefacts depends on shifting policy interpretations and enforcement mechanisms [Gil18]. Content moderation on large platforms is applied unevenly and is prone to error and bias [Gil18]. To address linguistic and regional variation, platforms report relying on language- and region-specific expertise when reviewing flagged content [Gil18].

The opacity of these governance regimes reinforces the transparency problem formalised later in Section 2.6, but in the present section it is treated primarily as an institutional condition of digital remembrance.

Socio-cultural conditions. Socio-cultural conditions also play a critical role [Ass11]. Remembrance is supported through practices of circulation and engagement, including sharing and participatory interaction [Gil18]. Digital artefacts that are not actively engaged with risk becoming less visible even when they remain technically accessible [Gil18]. Attention dynamics and audience interest therefore shape which materials remain salient over time. Because such engagement is selective and finite, some recorded materials recede into an archival background rather than remaining actively integrated into collective memory [Ass11].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Temporal conditions. Finally, temporal conditions underscore the fact that remembrance is a process rather than a static state [Ass11]. Over time, shifting platform policies and enforcement practices can affect what content is flagged and acted upon [Gil18]. More broadly, changes in storage systems, media conditions, and social organisation shape what can be retained and reactivated as memory [Ass11]. As a result, remembrance in digital environments must be understood as contingent and reversible, rather than permanent [Ass11, WH25].

Taken together, these conditions indicate that digital remembrance is structurally fragile and externally constrained. While they shape whether content can persist as part of collective memory, they are not directly observable in their entirety [Gil18, WH25]. Instead, their effects become visible only indirectly, through changes in accessibility, stability, and continuity over time. This conceptualisation establishes the theoretical boundary within which empirical observation must later operate.

2.5 Processes of Forgetting

Forgetting in digital environments is described not as a singular event of erasure, but as a set of processes through which the past becomes hidden or weakened and through which traces may remain available while becoming inaccessible (for example, difficult to find) [WH25]. Hoskins frames this environment as a forgetting ecology in which the relationship between media and memory is reversed "from . . . strengthening to hiding, dissipating, and weakening the past" [WH25]. The following paragraphs set out five premises that clarify this processual account and specify how it is used as an analytical framing in this thesis.

Forgetting as inaccessibility rather than erasure. Hoskins distinguishes between unavailability through total erasure and more common forms of forgetting that arise through inaccessibility, that is, content becoming difficult to find [WH25]. In this thesis, this distinction supports treating observed loss not only as deletion but also as practical forms of non-retrievability.

Mechanism plurality and partial persistence. Hoskins frames a principal form of contemporary forgetting as blockage: content may be available to memory while very little is humanly accessible [WH25]. This premise motivates treating forgetting as compatible with partial persistence and continuing traces, rather than equating it with absence.

Algorithmic mediation within a data-saturated ecology. Hoskins situates remembering and forgetting within an ecology characterised by the "messiness" of multiple connectivities and increasing saturation by data that mixes the human and the nonhuman [WH25]. He further frames the contemporary past as increasingly shaped by an algorithmic present that is difficult to capture and preserve for future use [WH25]. In

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

this thesis, this premise motivates an analytical focus on externally observable transitions in availability and retrievability without presupposing intentional erasure.

Temporal uncertainty and future contingency. Hoskins emphasises uncertainty over what content might endure to haunt the future or disappear without trace [WH25]. He also highlights that what is accessible at one moment may not remain accessible, raising the question of what will be accessible or not tomorrow [WH25]. This premise supports treating forgetting as temporally extended and contingent rather than as a single event.

Non-universality across ecologies and the social survival of memory. Hoskins notes that forgetting dynamics are not uniformly applicable and that not all (media and forgetting) ecologies are the same [WH25]. He further links the social existence of memories—their spread, support, and survival—to validation or direction by others, groups, and institutions [WH25]. In this thesis, this premise supports treating vulnerability to loss as context-dependent without specifying a universal mechanism.

This framing does not require generalisation across forgetting ecologies: Hoskins notes that it becomes impossible to generalise and that digital feeds shard highly individualised worlds [WH25]. Rather, it provides an analytic lens for describing how digital artefacts may transition through different states of accessibility over time. This perspective foregrounds blockage, sharding, and uncertainty over future accessibility, and highlights the need for empirical approaches that can accommodate partial and ambiguous forms of loss without presupposing causality or intent [WH25].

2.6 Digital Humanism as Research Foundation

Digital Humanism has emerged as an interdisciplinary initiative concerned with the societal consequences of digital transformation and the changing relationship between humans and digital technologies. Rather than constituting a domain-specific theory or an empirical framework, Digital Humanism articulates high-level premises concerning how digital systems should be situated within social reality and considered in relation to collective human concerns. In this thesis, Digital Humanism serves as a research motivational foundation: it motivates the investigation of platform-mediated digital memory, shapes the constraints and design decisions of the empirical analysis, and situates its findings within a principled normative context — without serving as an evaluative framework for the results themselves. The following paragraphs present five premises that define the problem space and specify the empirical scope of the study, connecting each to the specific principles and foundational claims of the Vienna Manifesto on Digital Humanism [WLA+19].

Non-neutrality of digital technology. A central premise of Digital Humanism is the rejection of technological neutrality. On this view, digital technologies and platforms

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

are treated as socio-technical systems that reshape social reality and contribute to transformations of key societal concepts [WPLG22]. This premise is grounded in the Manifesto's Principle 9, which holds that no technology is neutral and that practitioners must acknowledge their shared responsibility for the impact of information technologies, and in Foundational Claim 3, which observes that digital technologies are shaped by implicit and explicit choices that incorporate values, norms, and economic interests, many of which remain hidden in software programs implementing algorithms that remain invisible [WLA+19]. In this thesis, these premises motivate an analysis of content availability, visibility, and persistence as socially meaningful phenomena rather than technical side-effects of platform operation. They also delimit the methodological scope of the study: because platform-internal decision processes remain invisible to external observers, the empirical analysis is restricted to externally observable outcomes and does not infer hidden mechanisms, intentions, or causal rationales from those outcomes alone.

The 2017 large-scale removal of Syria-related YouTube documentation is significant for the theoretical framework developed in this thesis for two reasons. First, it instantiates at a documented, empirical scale the processual character of digital forgetting described by Hoskins: the loss was not a single erasure event but the downstream consequence of ongoing infrastructural changes to the platform's governance regime, operating through automated systems whose decision logic was not externally observable and whose effects accumulated over time [WH25]. Second, it illustrates the foundational claim of the Vienna Manifesto on Digital Humanism that digital technologies are shaped by implicit and explicit choices incorporating values, norms, and economic interests — and that many of these choices remain hidden in software programs implementing algorithms that remain invisible [WLA+19]. The Manifesto's Principle 9, which holds that no technology is neutral and that practitioners must acknowledge their shared responsibility for the consequences of information technologies, is directly relevant here, not as an evaluative claim about the conduct of any specific platform but as a conceptual framing that motivates the research problem [WLA+19]. The question of how to observe, document, and measure such consequences from an external position — without access to platform-internal decision-making processes — is precisely the methodological challenge this thesis addresses.

Societal-scale relevance. Closely related to this is the emphasis placed by Digital Humanism on societal-scale relevance. Digital systems are described as affecting not only individual users, but also collective practices, shared meanings, and socially consequential processes [Wer23]. This premise is grounded in the Manifesto's Foundational Claim 1, which identifies the disruption of societal institutions and political structures by the flood of data and algorithms as the defining challenge of contemporary digital transformation, and in Principle 2, which identifies the safeguarding of information dissemination as a central concern of platform governance [WLA+19]. This provides a conceptual basis for treating platform-hosted content as part of a broader environment in which collective memory is formed, maintained, or disrupted [Wer23]. Accordingly, the focus on Syria-related content on YouTube is motivated not by individual consumption patterns, but by

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the potential role of platform-mediated infrastructures in shaping historically relevant digital traces whose accessibility matters not only to current users but to the future communities that will seek to understand, contest, and build upon the documentary record of the Syrian conflict.

The Manifesto's Principle 2, which holds that platforms need to safeguard the free expression of opinion and the dissemination of information, similarly provides a framing context for the research problem without constituting an evaluative verdict on any specific set of platform decisions [WLA+19]. When the dissemination of information that constitutes a primary historical record of a major armed conflict is subject to the patterns of availability loss, archival fragmentation, and semantic degradation documented in this thesis, the question of how to observe those patterns reproducibly and transparently becomes a matter of societal significance rather than merely methodological interest.

This convergence between Digital Humanism's emphasis on societal-scale relevance and the memory-studies account of cultural memory is not merely thematic but methodologically significant. Assmann's framework establishes that cultural memory depends upon institutionalised carriers — material objects, practices, and institutions that enable the transmission of interpretations of the past across generations [Ass11]. In digital environments, platform-hosted content increasingly performs the function of such carriers, particularly for events that are sparsely documented through conventional institutional channels. When that content is subject to availability loss, the consequence is not merely the disappearance of individual records but the degradation of the infrastructural conditions under which collective memory can be formed, maintained, and transmitted. The societal-scale relevance invoked by Digital Humanism [Wer23] thus operates at the same level of analysis as the cultural memory theory that motivates the study of digital forgetting: both frameworks treat what is at stake in platform-mediated content loss as a matter of collective rather than merely individual significance. This thesis operates at the intersection of these frameworks, treating the availability and persistence of Syria-related YouTube documentation as a site where the societal consequences of platform governance become empirically observable.

Principle 6 holds that technological disciplines such as computer science/informatics must collaborate with social sciences, humanities, and other sciences rather than operating within disciplinary silos [WLA+19]. The observational measurement system developed here is constitutively interdisciplinary: it draws on web archiving research to specify the structure of archival evidence, on information systems research to ground the measurement framework, on memory studies to motivate the analytical categories, and on platform governance scholarship to delimit what can be claimed from observable signals.

Platforms as structural environments. Digital Humanism further conceptualises platforms as structural environments within which social processes unfold, rather than as isolated technical tools [Wer23]. This perspective foregrounds systemic conditions and infrastructural configurations over discrete events or individual actions. It is grounded in Foundational Claim 3's observation that the choices embedded in digital systems carry

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

structural consequences for the social world — consequences that cannot be reduced to the intentions of any individual actor — and in Principle 3's demand that the transparency and accountability of software programs and algorithms be ensured through effective regulation and research practice [WLA+19]. In the present thesis, this supports an analytical focus on longitudinal availability trajectories and cross-layer observability patterns rather than on isolated moderation decisions or explanations tied to individual platforms. Platforms are thus treated as structural conditions for the persistence and accessibility of digital traces. The opacity of these governance regimes — their resistance to external review and independent verification — is itself a condition of remembrance in the sense that the Vienna Manifesto on Digital Humanism identifies as requiring structural attention: the accountability and transparency of software programs and algorithms that the Manifesto calls for as a normative requirement would, if achieved, transform the conditions under which digital remembrance can be studied [WLA+19]. In the absence of such accountability, external observation of the kind developed in this thesis constitutes a necessary means by which the effects of these governance regimes on digital memory can be documented. The measurement framework responds to Principle 3's transparency requirement not by evaluating platform governance but by constructing a research infrastructure whose own derivation logic is fully explicit, independently verifiable, and auditable — making transparency an operational property of the research process rather than an aspiration applied to the platforms being studied.

Human-centred normative orientation. In addition to these descriptive premises, Digital Humanism articulates a normative concern for the human and societal consequences of digital transformation. Digital systems are analysed in this literature with reference to their implications for human interests, values, and social life [WPLG22, NRW22]. This orientation is expressed most directly in the Manifesto's Principle 5, which holds that decisions with consequences for individual or collective human rights must continue to be made by humans rather than automated systems, and that automated decision-making should support rather than replace human judgment [WLA+19]. In this thesis, that orientation functions as a motivational background for the research problem without being operationalised as an evaluative criterion for the empirical results or used to derive policy recommendations. What it does do is treat the consequences of automated systems for the observable record of human events as a legitimate and important object of empirical study — and design its measurement framework so that those consequences can be documented, bounded, and made transparent without exceeding what external observation can support. The practical contribution of the thesis — the reproducible observational measurement system — is the research infrastructure through which the transparent study of these consequences becomes possible.

Human-located moral responsibility. A further philosophical commitment associated with Digital Humanism is the attribution of moral responsibility and judgement exclusively to human agents, rather than to technical systems [NRW22]. This commitment supports a separation between empirical observation of platform-mediated outcomes and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

subsequent normative questions, and is grounded in the Manifesto's Principle 5 and in its broader concern with maintaining human accountability in an era of increasing automation [WLA+19]. Accordingly, the empirical analysis does not infer intent or responsibility from observed availability patterns and does not attribute moral agency to platforms. The patterns of right-censoring, cross-source non-alignment, and semantic loss documented in Chapters 4 and 5 are treated as properties of the measurement environment rather than as indicators of culpable conduct. This is not merely a methodological convention but a principled commitment: the Manifesto's insistence on human-located moral responsibility is respected by ensuring that the empirical analysis does not short-circuit the normative questions that Digital Humanism raises by pretending to answer them through statistical measurement. The RQ3 artefact — the graded completeness representation introduced in Chapter 5 — is designed to support rather than replace the human judgments that the constituent community must make about what to preserve, what to prioritise, and how to respond to the conditions of observability that the thesis documents.

Taken together, these five premises locate Digital Humanism in this thesis as an orienting context for why platform-mediated digital memory matters socially and why the conditions under which it can be empirically studied require explicit methodological attention. The Manifesto's foundational claims provide the diagnostic framework — invisible algorithms, co-evolution of technology and society, the quest for human-centred innovation — and its specific principles provide the normative orientation: transparency, accountability, interdisciplinarity, societal engagement, and human judgment as the irreducible locus of responsibility. The empirical chapters that follow focus on the conditions under which digital traces remain observable and reconstructable, with particular attention to the observability limits and measurement practices that shape what can be established from external evidence. The completeness artefact introduced in Chapter 3 is the most concrete instantiation of these principles within the research infrastructure: it makes the graded conditions of archival observability visible, comparable, and analytically tractable — not as an evaluation of platform governance but as a transparency instrument that enables the constituent community to understand the conditions under which the digital record of platform-mediated events can be studied, preserved, and contested.

In this thesis, the expression "Digital Humanism Mandate" is used in a bounded sense: Digital Humanism provides the motivational foundation for transparent, externally observable, and reproducible study of platform-mediated digital memory. Empirical explanation and evaluation are carried by the observational measurement framework.

2.6.1 Digital Humanism Across the Thesis

Table 2.1 maps each claim of the Vienna Manifesto on Digital Humanism [WLA+19] that is relevant to this thesis to the specific element through which it is operationalised, organised by the function the claim performs in the research.

Five Manifesto claims — P1, P4, P7, P10, and P11 — concern antitrust regulation, university policy, and educational curricula and fall outside the empirical scope of this

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 2.1: Overview of how the present thesis engages with the claims of the Vienna Manifesto on Digital Humanism [WLA+19]. Each cell names where the claim frames, constrains, informs, or is discussed within the thesis; the section mark (§) links to where the alignment is documented. Claim identifiers link to their glossary definitions. FC = Foundational Claim; P = Principle.

Claim	*Frames*	*Constrains*	*Designs*	*Grounds*
FC1	Syrian corpus as societal-scale object of study (§2.6)	—	—	—
FC3	Platform opacity as research problem (§2.6)	External-only observational design (§2.6)	—	Strictly observational analytical scope (§4.1)
P2	Syrian documentation as information dissemination at risk (§2.6)	—	—	RQ2 differential availability finding (§6.5)
P3	—	—	Provenance, Reference, Instrument, Method, Actor, Data (PRIMAD) reproducibility framework (§3.7); completeness artefact (§3.5.7)	—
P5	—	Non-evaluative observational scope (§2.6)	Non-interventionist artefact design (§3.5.7)	E2 infrastructure opacity finding (§6.2.2)
P6	—	—	Five-cluster Rigor Cycle grounding (§3.4); constituent community (§3.3.1)	Interdisciplinary measurement framework (§2.6)
P8	—	—	DSR outward flow; field testing obligation (§3.3.4); PRIMAD reproducibility framework (§3.7)	—
P9	Non-neutrality of platform governance (§2.6)	Rejection of causal attribution (§2.6)	—	—

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

thesis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 3

Methodology

3.1 Introduction

This thesis investigates the external observability of conflict-related YouTube content disappearance and its implications for the study of platform-mediated collective memory under conditions of platform opacity and infrastructural instability. Building on the theoretical framework developed in Chapter 2, the methodological approach adopts a system-oriented, observational research design that treats digital remembrance as conditional and digital forgetting as a processual, non-binary phenomenon. The central methodological challenge addressed in this chapter is not to explain why specific content disappears, but to document how changes in accessibility, visibility, and archival persistence can be observed over time through externally available signals.

The study adopts Design Science Research (DSR) because it combines analytical inquiry with the construction of reproducible research artefacts that enable transparent observation of platform- and archive-mediated memory dynamics [PTRC07]. Within this framework, the system developed in the thesis functions as a methodological instrument rather than a governance-level or policy-intervention tool [HC10]. It is designed to support longitudinal observation under conditions where platform-internal decision-making processes, moderation rationales, and policy enforcement mechanisms remain inaccessible to external researchers [HC10, GH13, PTRC07].

The research design does not attribute intent, causality, or responsibility to observed changes in content availability. Instead, it focuses on documenting externally observable states and transitions, while preserving ambiguity where evidence is inconclusive. This commitment reflects the Vienna Manifesto's Principle 9, which holds that no technology is neutral and that designers and deployers bear shared responsibility for the impact of digital systems: because platform governance and moderation mechanisms are not neutral instruments but value-laden choices embedded in infrastructure, this thesis declines to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

impute intent from observed outcomes and instead restricts analytical claims to what is externally and reproducibly observable.

The remainder of this chapter details how these commitments are operationalised through the Design Science Research framework. It describes the Relevance Cycle, the Design Cycle, and the Rigor Cycle, through which theoretical constructs are translated into data schemas, observational procedures, and reproducible artefacts. In doing so, the chapter establishes the methodological foundations for addressing **RQ1** (measurable patterns of content availability and archival persistence), **RQ2** (statistical associations with topic, language, and metadata), and **RQ3** (system-level design implications aligned with Digital Humanism). It also maintains a clear separation between empirical observation and interpretive analysis.

3.2 Research Framework: Design Science Research (DSR)

DSR provides a methodological framework for producing knowledge through the construction and evaluation of purposeful artefacts in response to defined problem contexts [HMPR04]. In this thesis, DSR structures the methodological approach through three interrelated cycles: the *Relevance Cycle*, the *Design Cycle*, and the *Rigor Cycle* [HMPR04, Hev07, GH13, MS95, PTRC07, Wie14]. Each cycle operates within the constraints established in Chapter 2, which delimit what can be claimed from external evidence under conditions of platform opacity.

The research gaps stated in Chapter 1 and the resulting research questions emerge from an iterative engagement with both the knowledge base and the empirical problem context. Following a hermeneutic understanding of literature review, the process of identifying relevant constructs and unresolved problems is not linear but develops through successive cycles of reading, interpretation, and refinement [BCK14]. Initial problem formulations derived from the Syria-related YouTube case and the constraints of platform opacity were progressively refined through engagement with the literature on digital memory, platform governance, and web archiving. In this sense, the gap formulation reflects the convergence of empirical problem recognition and iterative knowledge-base interpretation rather than the outcome of a predefined search procedure.

The *Relevance Cycle* connects the research to its application environment [Hev07]. It inputs requirements from the real-world problem context into the research and defines the acceptance criteria against which the utility of research artefacts is assessed. In this thesis, the Relevance Cycle is addressed in Section 3.3, which identifies the application domain and constituent community, evaluates the available data sources, defines the corpus sampling frame, and specifies the requirements and acceptance criteria that govern the Design Cycle. The application environment consists of the conditions under which platform-mediated digital memory can be studied externally: a corpus of conflict-related video content spanning 2011–2025, accessible only through platform-facing responses and web archive infrastructures. The constituent community addressed by this research com-

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

prises digital memory scholars, web archiving researchers, and civil society organisations engaged in the long-term preservation of conflict-related documentation.

The *Design Cycle* iterates between the construction and evaluation of design artefacts and processes [Hev07]. It is the central cycle of any DSR project, operating between the requirements supplied by the Relevance Cycle and the theoretical foundations supplied by the Rigor Cycle. In this thesis, the Design Cycle is addressed through two interconnected phases. The first phase covers the construction of the observational measurement system, including the semantic interpretation framework, the evidence-state classifier, the feature extraction layer, and the completeness construct (Section 3.5). The second phase covers the evaluation of those artefacts against methodological criteria derived from the Relevance Cycle (Section 3.6). The two phases are not sequential stages but constitute the iterative build and evaluate loop that defines the Design Cycle [HMPR04].

The *Rigor Cycle* connects the research to the knowledge base of scientific foundations, experience, and expertise that informs the research project, and adds the new knowledge generated by the research to that knowledge base [Hev07]. The inward flow of the Rigor Cycle is addressed by Chapter 2, which establishes the theoretical foundations that ground the design decisions executed in the Design Cycle: cultural memory theory [Ass11], scholarship on digital forgetting and platform governance [WH25], web archiving research [Bru18, BDA18], and measurement theory in Information Systems research [SBG04, Ste46]. The outward flow of the Rigor Cycle is addressed by Section 3.8, which documents the contributions that this research adds to the knowledge base.

This three-cycle structure aligns the theoretical commitments articulated in Chapter 2 with the system design and the stated methodological limits.

3.3 Relevance Cycle: Application Context and Problem Grounding

Methodological role of the Relevance Cycle. The Relevance Cycle requires methods capable of characterising the application environment, deriving requirements, and defining acceptance criteria [Hev07]. In this thesis, these functions are fulfilled through: (i) application domain analysis, (ii) constituent community characterisation, (iii) document and data source analysis, and (iv) corpus boundary definition. These activities collectively characterise the application environment and produce the requirements and acceptance criteria that flow into the Design Cycle, fulfilling the Relevance Cycle's inward flow function [Hev07].

3.3.1 Application Domain and Constituent Community

Application domain. The application domain of this thesis is the empirical study and preservation of platform-mediated digital memory under conditions of platform opacity. Since 2011, the Syrian conflict has produced one of the largest bodies of conflict-related audiovisual documentation on YouTube, generated by eyewitnesses, journalists, and civil

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

society actors [Syr25]. This material constitutes a primary form of digital documentation for events that are otherwise sparsely or partially documented in institutional archives. At the same time, this content is subject to ongoing loss through platform removal, access restriction, and archival degradation. These processes unfold gradually and often without external notification, under conditions of platform opacity that prevent independent observers from determining causes or verifying outcomes.

Constituent community. The constituent community of this research consists of digital memory scholars, web archiving researchers, and civil society organisations engaged in the long-term documentation and preservation of conflict-related audiovisual material. Organisations such as SyrianMemory systematically collect and curate platform-hosted conflict documentation, but lack reproducible, externally verifiable instruments for tracking what is being lost, at what rate, and under what observable conditions [Syr25].

The organisations addressed by this thesis operate within a historically specific and practically urgent context that extends beyond the methodological requirements of academic research. Civil society organisations engaged in the documentation and preservation of Syria-related audiovisual material — including the Syrian Archive (subsequently operating within Mnemonic), SyrianMemory, and affiliated digital memory initiatives — have developed their preservation practices in direct response to an ongoing and documented documentation crisis [Syr25, DH18, Ris19].

Operational problem context. This documentation crisis arises from the combination of three interacting conditions that define the operational environment within which these organisations work.

Condition 1: Platform opacity. The organisations that constitute the constituent community have no independent visibility into the moderation decisions, policy changes, or algorithmic systems that determine whether the content they seek to preserve remains available on the platforms that host it. The most consequential documented instance of this opacity's consequences was the 2017 deployment of automated content removal systems by YouTube, which resulted in the removal of over 400,000 conflict-related videos — more than ten per cent of all Syria-related content uploaded since 2011 — through hashing systems whose operational logic was not disclosed and whose effects were not preceded by notification [Hub22, AJAKD$^+$19, Tar24]. Civil society organisations working to preserve this material could not anticipate, monitor, or respond to such removals in real time [Hub22]. Platform opacity therefore defines not only an epistemological constraint on researchers, but also a structural vulnerability in the documentation environment within which these organisations operate.

Condition 2: Archival inconsistency. Web archiving infrastructures such as the Wayback Machine provide a partial and temporally uneven record of platform-hosted content, governed by crawl priorities and technical constraints that do not correspond to the evidentiary or historical significance of the material in question [Bru18, AAS$^+$11]. Prior work on web archive coverage documents that a substantial proportion of platform-hosted

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

content leaves no archival traces, and that among content with archival history, temporal gaps and semantic sparsity are prevalent [AAS+11, Bru18]. For civil society organisations seeking to document what has been lost and what remains accessible, the absence of comprehensive and reliable archival coverage means that their own preservation efforts cannot be systematically evaluated or compared against an independent baseline.

Condition 3: Ongoing documentation crisis. Unlike a completed historical event whose documentation has stabilised, the Syrian conflict and its aftermath continue to produce new documentation while simultaneously remaining subject to ongoing content loss through platform moderation, account terminations, and infrastructural change [Syr25, Whi24]. The constituent community operates within this environment in real time, making preservation decisions under conditions of incomplete information and without access to the tools needed to systematically observe what is being lost and at what rate.

Technological boundary conditions. The application environment is further constrained by the technical infrastructure available to the constituent community. Observation of platform-mediated content is limited to platform-facing interfaces, such as HTTP responses and publicly exposed metadata, and to web archiving infrastructures such as the Wayback Machine. The constituent community has no access to platform-internal data, moderation logs, or ranking systems, and lacks reproducible instruments for item-level longitudinal monitoring of content availability. These constraints define the technological boundary conditions under which the observational measurement system must operate.

Digital Humanism relevance. The application environment described above provides the local anchor for the Digital Humanism orientation established in Chapter 2. The Syria-related corpus is treated here as a societal-scale documentation environment rather than as a neutral technical dataset: its availability is shaped by platform and archival infrastructures whose internal operation cannot be fully observed from outside. This supports the relevance of Foundational Claim 1 without turning the Relevance Cycle into a separate theoretical discussion.

The constituent community also provides the local anchor for Principle 6. The organisations and research communities addressed by the thesis work across digital memory research, web archiving, software infrastructure, legal evidence preservation, journalism, and humanitarian documentation. The methodological implication is that the observational measurement system must be intelligible and useful across these communities rather than only within a single technical discipline.

Finally, Principle 8 is reflected in the decision to design an externally verifiable observational system. Because platform-internal data, moderation logs, and ranking systems are not available to independent observers, the thesis does not rely on inaccessible evidence or unsupported causal claims. Instead, it defines a measurement system whose inputs, transformations, and limits can be inspected by the constituent community.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Design implication. The methodological implication of this application environment is that the observational measurement system must combine heterogeneous external evidence while preserving the limits of each evidence source. It must integrate platform-facing responses, archival traces, and curated corpus metadata without treating them as equivalent forms of evidence. It must also support item-level longitudinal observation, retain partial and inconclusive traces, and document the boundary between what can be externally observed and what remains unknown. This design requirement flows from the Relevance Cycle into the Design Cycle and defines the empirical conditions under which the artefact is constructed.

3.3.2 Data Sources

Data source selection constitutes part of problem grounding in the Relevance Cycle because the available sources define both what can be observed and how it can be observed under conditions of platform opacity [HMPR04, Hev07]. All observation is constrained to publicly accessible platform interfaces and web archival infrastructures; no access to platform-internal data or moderation processes is available.

Table 3.1: Primary Data Sources and Their Analytical Purpose

Source	Purpose in the Thesis
SyrianMemory dataset [Syr25]	Baseline corpus of Syria-related YouTube videos, manually classified by NGO editors. Enables correlation of topical, temporal, and regional metadata with content availability.
YouTube (live HTTP endpoints)	Verification of current video availability and extraction of platform-exposed metadata fields (e.g., title, channel, publication date).
Wayback (CDX live HTTP)	Measurement of archival presence, capture continuity, and visibility over time to characterise patterns of digital forgetting.

Note. Bulk archival datasets such as *Common Crawl* [Com25] or precompiled *Internet Archive CDX dumps* [Int25a] were considered during the design phase. However, they were excluded because they provide static historical snapshots and do not allow direct observation of operational behaviours — such as rate limiting, selective accessibility, and temporal instability — that constitute the empirical object of this study. Working with the live Wayback infrastructure is therefore essential to examine digital amnesia as an evolving process rather than a fixed archive.

3.3.3 Corpus Definition and Sampling Strategy

The empirical scope of this study is defined by a corpus of approximately 800,000 Syria-related YouTube video identifiers collected from the SyrianMemory dataset and associated sources. From this corpus, a fixed analytical subset is constructed to support longitudinal observation of platform availability and archival traces using live web archiving infrastructures.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The analytical subset is defined through a stratified sampling procedure that explicitly specifies which video identifiers are included in the observational measurement system. This procedure constitutes a design decision within the Relevance Cycle of the DSR framework, as it establishes the empirical boundary conditions under which the observational artefact is instantiated and applied.

The resulting sample consists of approximately 50,000 unique video identifiers selected from the full corpus. Stratification is performed along two corpus-wide attributes available prior to archival processing: (i) temporal position based on publication date, aggregated at the level of year or year–month, and (ii) public availability status as recorded at the time of corpus construction. Video identifiers associated with missing or non-parseable publication dates are retained and assigned to a distinct *missing-date* stratum.

Within each stratum, video identifiers are ordered using a stable cryptographic hash function applied to the identifier value. A proportional quota is then allocated to each stratum based on its relative size in the full corpus, and identifiers are selected in hash order until the quota is met. Given the fixed hash-based ordering and proportional quota allocation, applying the procedure multiple times results in the same set of selected identifiers, irrespective of execution order or computational environment.

The sampled identifiers define the complete input set for all subsequent stages of the Design Cycle, including archival querying, semantic interpretation, trajectory reconstruction, and feature extraction. All empirical observations and quantitative analyses reported in later chapters are therefore conditioned on this explicitly defined sampling frame.

This procedure serves a dual role within the DSR framework. As an application environment boundary condition, it defines the empirical scope within which the observational artefact operates. As a constructed artefact, it is a reproducible design decision executed within the Design Cycle and documented as part of the measurement system [HMPR04, Hev07].

3.3.4 Requirements, Acceptance Criteria, and Outward Flow

The outputs of the activities above - the problem definition, the constituent community identification, the data source selection, and the corpus sampling frame - collectively constitute the *requirements* that flow from the Relevance Cycle into the Design Cycle [HMPR04, Hev07]. These requirements define four constraints the observational measurement system must satisfy, corresponding directly to the research gaps identified in Chapter 1: it must enable item-level longitudinal observation from external evidence only (**G1**); it must integrate heterogeneous and fragmented archival evidence layers without discarding partial traces (**G2**); it must represent availability as a graded, non-binary condition (**G3**); and it must produce reproducible, auditable artefacts that make the conditions and limits of observation explicit (**G4**).

The identified gaps (**G1**–**G4**) serve a dual role within the DSR framework. First, they define the research gap in the knowledge base and thus establish the contribution space

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

of the thesis. Second, they are translated into design requirements that constrain the Design Cycle. This dual role reflects the intersection of knowledge base limitations and application environment needs.

These constraints define the boundary conditions under which the Design Cycle constructs the observational artefact and the Rigor Cycle provides the conceptual and methodological grounding.

The *acceptance criterion* for the artefact is whether the observational measurement system produces reproducible, analytically tractable representations of platform-mediated digital memory that treat indeterminacy and partial archival traces as first-class outcomes, without exceeding what the external evidence supports. This criterion is what the Design Cycle evaluation in Section 3.6 is measured against.

Per the DSR framework, the outward flow of the Relevance Cycle requires that the artefact be returned to the constituent community for field testing in the application domain [Hev07]. In this thesis, that would mean deploying the observational measurement system with organisations such as SyrianMemory and Mnemonic to evaluate whether it improves their capacity to monitor platform-mediated content loss in practice. This step is not completed within the scope of this thesis and constitutes the primary direction for future work, implying that the Relevance Cycle does not fully close within the scope of this thesis [Hev07].

3.4 Rigor Cycle: Knowledge Base

3.4.1 Methodological function of the Rigor Cycle

The Rigor Cycle connects a DSR project to the existing knowledge base and translates prior constructs, models, and methods into design inputs for artefact development and evaluation [HMPR04, Hev07]. In this thesis, that inward flow is operationalised through a semi-systematic and integrative literature grounding procedure. This review logic is appropriate because the research problem spans multiple knowledge domains that have been conceptualised differently across disciplines, making a narrowly systematic evidence synthesis neither feasible nor suitable for the methodological purpose of this section [Sny19].

The purpose of the Rigor Cycle in this thesis is therefore not to produce a standalone review contribution about the literature itself. Instead, it translates existing knowledge into two forms of use in the thesis. *Design Cycle inputs:* a construct dictionary, a variable list, and a set of documented methodological inputs that ground the observational measurement system. *Thesis-wide grounding uses:* the same grounded literature is used in later chapters to set the scope of the analysis, limit what can be inferred from external evidence, interpret the findings, and support the methodological, empirical, and conceptual contributions of the thesis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.4.2 Literature grounding procedure

Purpose and scope

The inward flow of the Rigor Cycle was implemented through a documented literature grounding procedure whose purpose was to connect existing knowledge to the design and operationalisation of the observational pipeline [HMPR04, Hev07]. The procedure was designed to cover five thematic clusters relevant to the thesis:

- digital memory and digital amnesia
- collective memory
- platform governance
- web archiving
- measurement, reproducibility, and design science

The review covered the period 2011–2025. This boundary was selected to match the temporal context of the Syrian conflict and the digital documentation environment within which the empirical application of the thesis is situated. All identified sources were managed in Zotero under the collection *Literature_Grounding_Corpus*.

Search and source identification

Sources were identified through database searches in ACM Digital Library, IEEE Xplore, Scopus, and Google Scholar, complemented by key seed texts and backward and forward citation tracking. The search strategy was semi-systematic in the sense described by Snyder: it was designed to cover a broad and conceptually heterogeneous field without claiming exhaustive retrieval of all possibly relevant publications [Sny19]. At the same time, the procedure also followed an integrative logic because the aim was to bring concepts and procedures from different knowledge domains into a single operational framework for the design process [Sny19].

Search, reading, and refinement were iterative throughout. Reading identified additional search terms, conceptual distinctions, and citation paths, which in turn informed further retrieval. This iterative character is consistent with the view that literature reviewing develops through repeated movement between searching, reading, interpretation, and synthesis rather than through a strictly linear sequence [BCK14].

Selection and classification

Sources were included when they contributed at least one of the following: a concept definition, an operationalisation, an observable indicator, a methodological distinction, or a procedural specification relevant to one of the five thematic clusters. Sources that

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

provided only high-level theoretical orientation without operationalisable content were used for conceptual framing in Chapter 2, but were not entered into the construct dictionary used for methodological grounding.

Each included source was then classified in two ways. First, it was assigned to one of the five thematic clusters. Second, it was labelled according to its primary role in the thesis as either conceptual or methodological. This distinction supported later separation between substantive constructs and procedural precedents.

Extraction, synthesis, and operationalisation

For each included source, relevant material was extracted, including definitions, conceptual distinctions, observable indicators, and procedural specifications. Extracted elements were then compared across sources and grouped where they expressed the same or closely related analytical ideas. These grouped elements were subsequently defined as shared constructs for use in this thesis.

Where required, the resulting constructs were translated into operational definitions and then into measurable variables suitable for the observational pipeline. For transparency, the analytical steps can be reported as follows:

1. **Selection:** identify sources that contribute operationally relevant conceptual or methodological content.
2. **Classification:** assign each source to a thematic cluster and to a conceptual or methodological role.
3. **Extraction:** record relevant concepts, distinctions, indicators, and procedures.
4. **Synthesis:** compare and group extracted elements across sources to define shared constructs.
5. **Operationalisation:** translate constructs into operational definitions and measurable variables where required.
6. **Application:** implement the resulting variables and methodological precedents in the observational pipeline and document them in the appendix tables.

These steps are reported sequentially for auditability. However, the overall review process was iterative rather than linear: searching, reading, comparison, and operationalisation repeatedly informed one another throughout the grounding procedure [BCK14].

This procedure establishes a direct trace from literature to implementation:

$$\text{source} \rightarrow \text{concept} \rightarrow \text{construct} \rightarrow \text{variable} \rightarrow \text{analysis}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

For example, the concept of content availability is operationalised into $avail_v^{\text{html}}$ and $html_state_i$; temporal discontinuity in archival records is operationalised into δ_v; and the onset of observed availability states is represented by $t_{\text{avail_start}}$ and t_{disp}.

Stopping rule, limitation, and documented outputs

The literature grounding procedure was considered complete when additional sources no longer generated new constructs, distinctions, or methodological procedures relevant to the five thematic clusters [BCK14]. In this sense, completion was determined by pragmatic saturation of methodological yield rather than by a claim of exhaustive coverage.

A limitation of this procedure is that source selection, extraction, and grouping were conducted by a single researcher. As a result, the saturation judgement and the consolidation of constructs remain interpretive and may have differed under a multi-reviewer procedure.

The outputs of the grounding procedure are documented in Appendix B.1:

- Table B.1 lists the constructs and their definitions, distinguishing operationalised constructs from those used only for theoretical grounding.
- Table B.2 lists the variables derived from these constructs and used in the analysis.
- Table B.3 lists the methodological precedents drawn from the literature and applied in data collection, processing, and evaluation.

3.4.3 Knowledge base inputs to the Design Cycle

The literature grounding procedure identified an overarching research foundation and five more immediate bodies of knowledge that informed the Design Cycle. Together, these inputs provide the conceptual and methodological basis for the observational system developed in this thesis.

Digital Humanism as overarching research motivational foundation. Digital Humanism, as established in Chapter 2, provides the overarching research motivational foundation of this thesis and situates the Design Cycle within that broader frame. In methodological terms, it frames platform-mediated digital memory as a socially significant problem, limits the analysis to externally observable evidence, and supports a transparent, reproducible, interdisciplinary, and non-interventionist design of the observational artefact. In this sense, Digital Humanism does not function here as a domain-specific explanatory theory, but as a higher-level orientation for what is studied, what can be claimed from the evidence, and how the research infrastructure is designed.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Substantive knowledge domains

Cultural and collective memory theory. Cultural and collective memory scholarship [Ass11] establishes that remembrance depends on technical, institutional, and infrastructural conditions rather than following automatically from the existence of content. In this thesis, this grounding constrains the semantic interpretation of archival traces: observed records are treated as indicators of preservation conditions, not as direct proxies for continuous content existence.

Digital forgetting and platform governance. Scholarship on digital forgetting and platform governance [WH25, Gil18] supports the treatment of availability loss as processual, non-binary, and frequently opaque to external observers. This grounding informs the *7+1 archive-side evidence-state model* and justifies retaining indeterminate states as explicit analytical outcomes rather than collapsing them into missing data.

Web archiving research. Web archiving scholarship [Bru18, BDA18, AAS+11] establishes that archival records are incomplete, temporally uneven, and not semantically identical to live-platform states. This grounding motivates the two-layer CDX/HTML classifier design and the explicit treatment of cross-source non-alignment in Section 3.5.4.

Measurement and procedural knowledge domains

Measurement theory in Information Systems. Measurement theory in Information Systems research [Ste46, SBG04] provides the basis for transforming observable system artefacts into analytically defensible constructs. In this thesis, it structures the operationalisation of completeness in Section 3.5.7, including the ordinal encoding logic, the bounding rule, and the distinction between modal preservation state and secondary quantitative modifiers.

Methodological frameworks. Established methodological frameworks support the procedural organisation and reproducibility of the Design Cycle. CRISP-DM provides a structured basis for timeline integration, data preparation, modelling, and logging [WH00]. PRIMAD provides the basis for documenting provenance, instrument versions, methods, and data sources across runs [FFJ+16]. These frameworks therefore contribute procedural discipline rather than substantive constructs.

Taken together, these five bodies of knowledge do not function as parallel references added to the chapter for completeness. They are integrated as design inputs that constrain interpretation, guide operationalisation, and structure reproducible execution within the Design Cycle. In this sense, the inward flow of the Rigor Cycle provides a traceable bridge between the knowledge base and artefact construction.

3.4.4 Outward flow: contributions to the knowledge base

The outward flow of the Rigor Cycle identifies the kinds of knowledge that the thesis contributes back to the DSR knowledge base [HMPR04, Hev07]. In this thesis, these contributions are classified as constructs, a model, methods, an instantiation, and a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

methodology contribution. This subsection records the contribution categories in DSR terms; the broader empirical, methodological, and practical contribution statement is presented in Section 3.8.

Constructs. The thesis contributes two constructs: the graded archival observability representation operationalised through the archival completeness construct (Section 3.5.7), and the *7+1 archive-side evidence-state model* used for longitudinal platform-mediated content observation (Section 3.5.4).

Model. The finite-state machine formulation of the evidence-state classifier constitutes the model contribution. It formalises the transition logic over the *7+1 archive-side evidence-state model* and makes the classifier structure explicit and inspectable (Section 3.5.4).

Methods. The thesis contributes three procedural methods: the two-layered CDX/HTML classification procedure (Section 3.5.4), the WARC-first fallback parsing strategy (Section 3.5.8), and the stratified cryptographic-hash sampling protocol (Section 3.3.3).

Instantiation. The implemented observational measurement system, versioned as `v1.1-thesis`, constitutes the instantiation contribution. It includes the LoadBench infrastructure, Wayback Module, YouTube Module, SQLite data store, derivation layer, and feature construction layer [Alk26].

Methodology. The PRIMAD-instrumented reproducibility procedure constitutes the methodology contribution. It documents provenance, instrument versions, methods, data sources, and validation outputs across repeated runs, thereby supporting reproducible longitudinal platform observability research (Section 3.6).

This outward-flow classification is used only to locate the thesis contributions within the DSR knowledge base. Their substantive scope, limitations, and relation to the research questions are discussed in Section 3.8.

3.5 Design Cycle: Artefact Design

3.5.1 Components of the Observational Measurement System

This section specifies the construction phase of the *Design Cycle* by presenting the observational measurement system developed in this thesis. All system components described in this section operate exclusively on the fixed analytical subset of video identifiers defined by the corpus sampling strategy in Section 3.3.3. The system comprises a sequence of interdependent artefacts that translate externally observable archival and platform-facing signals into analytically usable representations, without invoking platform-internal knowledge or causal inference.

The construction phase comprises the following system components:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Component M1: Semantic Interpretation Framework. Defines how raw signals from the Internet Archive layer and the YouTube platform layer are interpreted. This component specifies the observation layers and the interpretation rule logic that maps observed signals to semantic state variables. It is specified in Section 3.5.3.

Component M2: Archival Availability Trajectory Reconstruction Constructs time-indexed availability trajectories from archival evidence. This component specifies the 7+1 archive-side evidence-state model, transition rules, and classification algorithm, and produces evidence-state sequences, onset times, and event indicators for each video. It is specified in Section 3.5.4.

Component M3: Cross-Source Observability Boundary Assessment. Evaluates the relationship between archival evidence and live platform observations. This component determines whether and how signals from the Internet Archive layer and the YouTube platform layer can be compared, and produces indicators of cross-source observability and joint state distributions. It is specified in Section 3.5.5.

Component M4: Human Taxonomy Features Construct. Transforms curated taxonomy labels into structured analytical variables. This component defines the mapping, normalisation, and admissibility rules used to construct content-related features for subsequent analysis. It is specified in Section 3.5.6.

Component M5: Archival Completeness Construct. Operationalises the degree of archival preservation as a graded measure derived from observable archival layers. This component defines preservation states, associated scores, and aggregation rules used to quantify archival completeness. It is specified in Section 3.5.7.

Component M6: Quantitative Observational Analysis. Applies descriptive and statistical procedures to the resulting datasets. It is specified in Section 3.5.8 and produces the empirical results reported in Chapters 4 and 5.

Together, these components define a fully specified and repeatable procedure for examining platform-mediated digital memory under conditions of infrastructural instability and platform opacity.

Figure 3.1 shows the complete structure of the observational measurement system as it emerged from the iterative design process described above. The diagram illustrates how the empirical problem of unstable video availability and incomplete archival capture is translated into concrete technical requirements, including rule-based data processing, collection of observable payload evidence, and run-specific data storage. The system contains two observation modules: one for Wayback snapshots and one for YouTube platform states. Both modules access external endpoints through a shared infrastructure layer (LoadBench). All retrieved records are written to a single SQLite database, which

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

serves as the central data store for further processing. On the basis of the stored records, the system calculates archival completeness scores and reconstructs availability trajectories. These outputs are used to generate the feature matrices for RQ1 and RQ2, which form the exclusive input for the statistical analyses reported in the following chapters.

The presentation of the Relevance, Design, and Rigor Cycles follows a sequential structure for clarity. In practice, these cycles operated iteratively and interactively, with design decisions shaped by simultaneous inputs from both the application environment and the knowledge base [Hev07].

3.5.2 Iterative Construction and Evaluation Within the Design Cycle

The *Design Cycle* in DSR iterates between the construction of artefacts and their evaluation, with each evaluation producing feedback that either confirms the current design or identifies a requirement for revision [HMPR04, Hev07]. Multiple iterations are completed before artefact specifications are stabilised and contributions are released to the knowledge base or the application environment [Hev07]. The artefact specifications presented in Sections 3.5.3–3.5.8 are the result of this iterative process. Two evaluation outcomes produced documented design revisions.

First, the evidence-state classifier was initially constructed using CDX index records as the sole input source. Evaluation of this approach established that CDX metadata provides only coarse temporal structure and does not resolve availability states at the page level. As established in prior web-archiving scholarship, CDX records document the interaction between the archival crawler and the target resource rather than the semantic availability of the resource itself [AAS+11]. A CDX-only classifier therefore cannot distinguish between a video that was accessible at a given timestamp and one that returned a structurally valid HTTP response without exposing platform-level availability semantics. This evaluation finding introduced a requirement that the CDX-only design could not satisfy. The classifier was revised to a two-layered design: a CDX-only stage that provides corpus-wide temporal structure, and an HTML-enhanced stage that refines evidence-state transitions where replayable HTML captures are available. The revised design applies HTML-based refinement only where such payloads exist, leaving videos without replayable HTML classified solely via archival interaction signals. This prevents the system from producing interpretations not supported by the available observations, as documented in Section 3.5.4.

Second, the binary availability classification adopted in prior empirical studies of platform-hosted content [Gil18, GBK20] was evaluated against the observational requirements of this thesis. Binary classification treats content as either present or removed, which is analytically insufficient for a corpus in which content may lose contextual metadata without becoming unavailable, remain indexed without being replayable, or transition through intermediate access-restriction states before reaching a terminal unavailability state. These intermediate conditions constitute empirically distinct and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

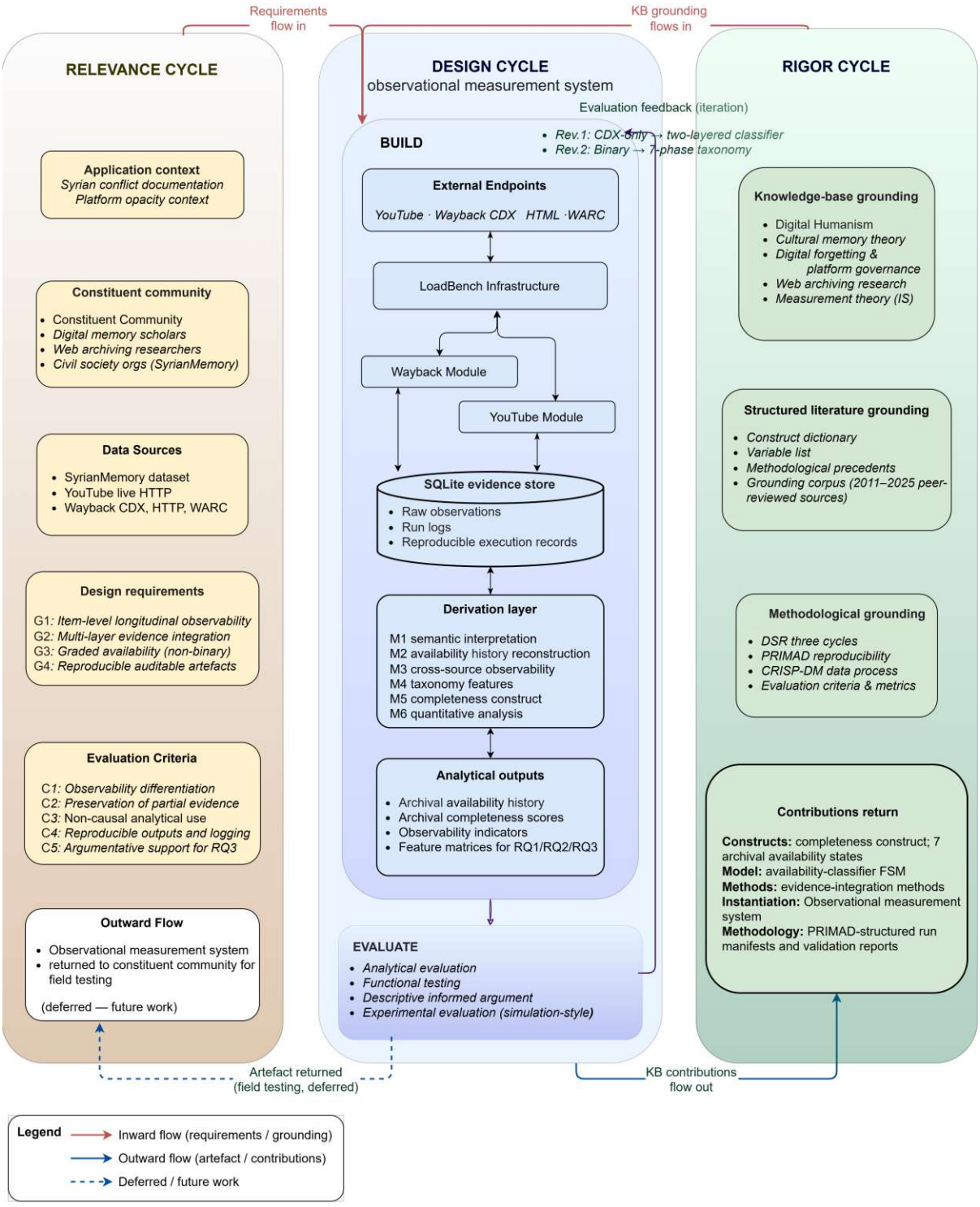

Figure 3.1: DSR-architecture of the observational measurement system.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

analytically separable states, not noise or missing data. The model was therefore extended from a binary availability outcome to archive-side evidence states — Available, Restricted, Eroded, Unavailable, Redirect-only, Archival gap, Terminal absence, and residual Unknown archive evidence — each defined by a distinct and mutually interpretable configuration of externally observable signals. The full specification is provided in Section 3.5.4.

The artefact specifications presented in the following sections are the stabilised output of these iterations. Each specification was retained after evaluation confirmed that it extended the observational range of the system without introducing inference beyond what the available archival evidence supports.

3.5.3 Component M1: Semantic Interpretation Framework

The semantic interpretation framework defines how availability-related external evidence is converted into state variables used by the subsequent components of the observational measurement system. It is part of a broader framework that combines heterogeneous external evidence, including archival traces from the Wayback Machine, live platform-facing responses from YouTube, and curated corpus metadata. Within Component M1, the focus is limited to the two evidence sources that directly describe availability: the Internet Archive layer and the YouTube platform layer.

The *Internet Archive layer* provides historical archival evidence about past platform-facing states, derived from Wayback CDX snapshot metadata and archived HTML snapshots retrieved via the wayback machine replay layer. The *YouTube platform layer* provides a single point-in-time platform-facing availability observation per video, derived from live platform checks executed against YouTube endpoints.

The two availability-related layers are integrated within the same measurement system, but they are interpreted through source-specific rules. They are not collapsed into a single final availability status because they observe different infrastructures, different time points, and different classes of evidence. The Internet Archive layer supports longitudinal reconstruction, while the YouTube platform layer provides a current platform-facing reference point. Their relationship is examined explicitly in the cross-source observability boundary assessment specified in Section 3.5.5.

Purpose and scope

The reconstruction of digital remembrance and forgetting within platform–archive ecosystems requires an explicit semantic interpretation layer that translates heterogeneous archival and platform-facing signals into analytically meaningful observational states. In line with cultural memory theory, digital remembrance is treated as conditional upon technical, institutional, and infrastructural support rather than as an automatic consequence of content production [Ass11]. Digital forgetting is accordingly understood as a gradual, non-binary process shaped by socio-technical mediation rather than by singular acts of erasure [WH25].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The semantic interpretation framework developed in this thesis serves a methodological function. It structures externally observable evidence so that longitudinal patterns of memory persistence, degradation, and absence can be examined without reliance on platform-internal knowledge. The system is constructed to avoid producing knowledge claims that exceed what is externally observable under conditions of platform opacity. All interpreted signals are therefore treated as evidence of observable interfaces, archival interactions, and preservation conditions rather than as proxies for platform governance decisions.

The scope of the framework is explicitly descriptive and observational. It is designed to support (i) the identification of measurable longitudinal patterns of archival presence and absence (**RQ1**); (ii) comparative analysis of how different categories of content experience unequal conditions of digital remembrance (**RQ2**); and (iii) the identification of infrastructural limitations that motivate system-level design considerations consistent with Digital Humanism (**RQ3**) [NR21]. By maintaining this scope, the framework supports transparency, interpretability, and restraint in what is claimed under conditions of platform opacity.

Observation layer separation

The Internet Archive layer and the YouTube platform layer are separated because they observe different types of evidence. The Internet Archive layer reconstructs historical archival evidence from sequences of CDX records and replayable archived HTML snapshots. It provides evidence about how a video page was captured, indexed, replayed, or rendered by the archival infrastructure at different points in time.

The YouTube platform layer observes the current platform-facing state of each video identifier at the time of measurement. It does not reconstruct historical trajectories. Instead, it records the response returned by the live platform to a structured request and translates the externally visible endpoint and HTML signals into point-in-time availability variables.

This separation prevents historical archival evidence and current platform-facing evidence from being conflated. The Internet Archive layer provides the input for availability trajectory reconstruction in Section 3.5.4. The YouTube platform layer provides a contemporary reference point used in the empirical analysis and in the cross-source observability boundary assessment. Their relationship is analysed explicitly in Section 3.5.5.

Internet Archive layer

Inputs and archival interaction signals. Archival interaction signals constitute the primary empirical surface through which the behaviour of web archiving infrastructures can be observed over time. In this thesis, these signals are derived from the Internet Archive's CDX index, which records metadata for individual capture attempts, including timestamps, recorded HTTP response codes at crawl time, MIME types, and references to archived payloads stored in WARC files [Int25a, Int25b].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Prior web-archiving research emphasises that CDX records document the interaction between the archival crawler and the target resource rather than the semantic availability of the resource itself [AAS+11]. Consequently, CDX attributes are interpreted here as indicators of archival engagement and recording stability. Capture timestamps enable the reconstruction of longitudinal interaction sequences, while variations in capture frequency and continuity reveal periods of stable archival presence, intermittent interaction, or extended absence.

Status codes and MIME types recorded in CDX entries are treated strictly as technical descriptors of capture attempts, not as indicators of replayability or platform-side behaviour. This interpretation follows established documentation and scholarship on IA's indexing and replay mechanisms, which caution against equating CDX metadata with live-web semantics [AAS+11, HGA16]. Within the semantic interpretation framework, these attributes contribute to the identification of archival stability, fragmentation, and absence as infrastructural phenomena.

HTML-level availability and metadata signals. While CDX-based records document interactions between the archival crawler and the target resource, archived HTML snapshots provide access to the page content as it was displayed at the time of capture. For platforms such as YouTube, removal or loss of access is often expressed directly within the page content through textual notices embedded in the HTML, even when the underlying URL remains technically reachable. Previous studies have shown that archived YouTube pages can be used to identify instances of content unavailability based on such page-level messages preserved in web archives, rather than on distinctive HTTP response codes [Bru18, AAS+11].

Archived HTML representations enable the identification of observable availability states as they appeared to users at the time of capture. These include generic unavailability messages, access restrictions (e.g. private or region-restricted content), and other platform-rendered notices preserved in the archival record. The presence of such notices does not imply a singular underlying cause or enforcement mechanism; instead, it provides evidence of how content was presented within the platform interface at a specific point in time.

In addition to availability signals, web-archiving scholarship has noted that archived representations of dynamic video platforms frequently lack complete contextual metadata, such as titles, descriptions, or channel identifiers, limiting interpretability and evidentiary richness [Aas19]. Within the semantic interpretation framework of this thesis, metadata incompleteness is therefore treated as an observable property of archival representation rather than as evidence of a deletion decision or coordinated removal processes [Bru18, Aas19, KSBVdS19].

Temporal patterns of archival stability and fragmentation. Temporal organisation is central to the empirical observation of digital remembrance and forgetting. Rather than analysing isolated archival snapshots, the semantic interpretation framework

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

examines sequences of archival interaction and HTML-level signals in order to characterise recurring temporal patterns of stability, fragmentation, and absence [BDA18]. This perspective aligns with web-archiving scholarship that conceptualises archives as temporally uneven and structurally fragmented memory infrastructures, shaped by heterogeneous crawl frequencies, technical constraints, and institutional priorities [BDA18].

Within this thesis, archival stability is operationalised as periods of relatively regular capture activity accompanied by persistent contextual signals in the archival record. Fragmentation is observed when capture sequences become irregular, discontinuous, or intermittently absent, resulting in reduced temporal continuity. Extended temporal gaps are treated as a distinct observational pattern, representing structural breaks that limit longitudinal interpretability and reduce the evidentiary density of the archived record.

Taken together, these temporal patterns provide an empirical basis for examining how conditions of digital remembrance vary across time and content categories. By aggregating such patterns across large collections, the methodology enables comparative analysis of unequal archival persistence and makes differences associated with infrastructural constraints observable.

Interpretation rules and output states. For each archival interaction record

$$s_i = (t_i, \mathit{status}_i, \mathit{mime}_i),$$

the variables status_i and mime_i are treated as technical descriptors of crawler interaction and capture attempts, not as direct evidence of replayability or platform-side behaviour. When $\mathit{mime}_i = \texttt{text/html}$ and a replayable archived HTML payload is available, the archived representation is interpreted at the interface level and assigned a semantic annotation

$$\mathit{html_state}_i \in \{\mathit{available}, \mathit{restricted}, \mathit{unavailable}, \mathit{eroded}\}.$$

Within this interpretation:

- $\mathit{html_state}_i = \mathit{available}$ indicates that the archived page and embedded content were accessible in the archived representation.
- $\mathit{html_state}_i = \mathit{restricted}$ indicates interface-level access limitations such as age-gating, region restrictions, or authentication requirements.
- $\mathit{html_state}_i = \mathit{unavailable}$ indicates that the archived page displays a platform-level unavailability notice.
- $\mathit{html_state}_i = \mathit{eroded}$ indicates observable loss or degradation of contextual metadata relative to earlier captures.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Metadata incompleteness is treated as a property of archival representation rather than as evidence of deletion. Extended temporal gaps between *cdx* observations are treated as observational discontinuities rather than as automatic evidence of disappearance.

The semantic interpretation framework integrates archival interaction signals, HTML-level availability and metadata signals, and temporal patterns into a conceptual classification of digital remembrance states. This classification provides a structured vocabulary for describing empirically observable conditions of archival persistence, degradation, and absence as they appear in archival records over time. It serves as the interpretive layer of the measurement system. The operational 7+1 archive-side evidence-state model used by the reconstruction algorithm is specified separately in Section 3.5.4.

Each remembrance state corresponds to a recurring configuration of externally observable signals. Stable archival remembrance denotes periods in which replayable HTML snapshots occur at relatively regular intervals and contextual metadata (e.g. title, description, channel information) is present in those captures. Fragile or degraded remembrance describes situations in which the archive still contains captures of the video page, but capture intervals are irregular or include gaps exceeding the predefined threshold, or contextual metadata is partially missing. Archival absence refers to periods exceeding the predefined threshold Δt during which no replayable HTML captures are recorded in the archival record.

This classification is developed as a methodological artefact within the Design Science Research framework to support systematic comparison, aggregation, and longitudinal analysis of archival evidence [HC10]. It provides an analytical representation of observed archival conditions without asserting underlying platform behaviour and supports empirical investigation of memory dynamics across large collections of platform-mediated content [HC10].

YouTube platform layer

Purpose and inputs. The YouTube platform layer produces a point-in-time observation of the platform-facing availability state of each video identifier at the moment the measurement system is executed. Unlike the Internet Archive layer, which reconstructs historical availability states from sequences of archival snapshots, the YouTube platform layer observes the current state of each video as it is presented by the live platform at the time of measurement.

The layer is designed to operate without access to platform-internal decision-making processes, moderation rationales, or enforcement records. All observations are derived from externally accessible signals returned by the platform in response to a structured request. No causal inferences about platform behaviour, content removal, or moderation intent are drawn from these signals.

For each video identifier v, the YouTube platform layer executes a single live request to the YouTube platform and records the following inputs:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- **Endpoint signals**: the HTTP status code $status_v^{\mathrm{YT}}$ and the associated error classification $error_type_v$ derived from the platform's response metadata, together with the request execution timestamp t_{check}. These signals are derived from the HTTP response metadata alone, independently of the response body content, and encode the platform's disposition toward the request, including whether the video is accessible, the account is terminated, or the request was rate-limited.

- **HTML payload**: the raw HTML response body returned by the platform, when present. This payload may contain a structured player response embedded in the page that exposes the platform's internal playability assessment for the video. When no structured player response is present, for example when the platform returns a rate-limiting interstitial, the payload does not expose video-level availability semantics and is treated as analytically uninformative for availability classification purposes.

Two-signal classification design. The YouTube platform layer applies two independent classification procedures to the inputs above, producing two distinct output signals per video. These signals are stored separately and serve different analytical purposes.

Signal 1: Endpoint-derived availability. The first classification procedure operates exclusively on endpoint signals. It maps the observed combination of HTTP status code and error classification to a set of platform-facing availability categories that reflect the disposition of the request at the endpoint level. These categories include states such as *available*, *age-restricted*, *account-terminated*, *error*, and *unknown*. The procedure is deterministic: identical endpoint signals always produce identical category assignments. The resulting classification is used as the primary basis for corpus-level availability characterisation in Chapter 4 and as a stratifying variable in the cross-source observability boundary assessment (Section 3.5.5).

Signal 2: HTML-derived coarse availability. The second classification procedure operates on the HTML payload when present. It attempts to extract the structured player response embedded in the page and derives a coarse availability classification from the platform's own playability assessment contained within it. The procedure produces one of three outputs:

1. **Playable**: the structured player response is present and indicates that the video is accessible and playable on the platform.

2. **Explicitly unavailable**: the structured player response is present and indicates that the video is not playable, regardless of the specific reason. Alternatively, in the absence of a structured player response, the presence of known platform-rendered unavailability markers in the HTML payload produces this classification.

3. **Indeterminate**: the HTML payload is absent, not decodable, or does not contain either a structured player response or known unavailability markers. This outcome

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

indicates that the payload did not yield usable availability semantics and is treated as an unobservable state rather than as evidence of availability or unavailability.

The *indeterminate* outcome is the primary condition under which the HTML payload is treated as analytically uninformative. This occurs most systematically when the platform returns a rate-limiting response: the returned HTML consists of a platform-generated interstitial that does not contain a player response or availability markers, and the HTML-derived signal therefore yields *indeterminate*. In these cases, the endpoint-derived signal yields *unknown* for the same request, and both signals converge on unobservability.

The two signals are independent. A video may receive an endpoint-derived classification of *age-restricted* while yielding an HTML-derived classification of *explicitly unavailable* if the player response confirms inaccessibility, or *indeterminate* if the payload did not expose a player response. The two signals are never merged into a single unified classification; their relationship is examined in the cross-source observability boundary assessment (Section 3.5.5).

Outputs and analytical use. For each video identifier v, the YouTube platform layer produces:

$$Y(v) = (t_{\text{check}},\ avail_v^{\text{endpoint}},\ avail_v^{\text{html}}),$$

where t_{check} is the request execution timestamp, $avail_v^{\text{endpoint}}$ is the endpoint-derived availability category, and

$$avail_v^{\text{html}} \in \{playable,\ explicitly_unavailable,\ indeterminate\}$$

is the HTML-derived coarse availability classification.

Both output fields are stored in the central data store per video identifier and per run, and both are fixed at the time of execution. Neither field is updated or reprocessed after the run is complete.

The *endpoint-derived availability* field $avail_v^{\text{endpoint}}$ is used in Chapter 4 to characterise the distribution of platform-facing availability states across the corpus at the time of measurement, and in the cross-source observability boundary assessment to compare platform-facing states against prior archival HTML evidence (Section 3.5.5).

The *HTML-derived coarse availability* field $avail_v^{\text{html}}$ is used in Chapter 5 as the dependent variable for the **RQ2** regression analysis, where *playable* encodes $Y_i = 1$ and *explicitly_unavailable* encodes $Y_i = 0$. Videos yielding *indeterminate* are excluded from the **RQ2** analytical population because their availability status cannot be determined from the available evidence.

The YouTube platform layer output $Y(v)$ is not used in the availability trajectory reconstruction (Section 3.5.4), which operates exclusively on Internet Archive layer evidence. The point-in-time nature of the YouTube platform layer observation means it cannot independently support longitudinal analysis. Its contribution to the measurement

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

system is as a contemporary platform-facing reference point against which archival evidence is compared.

Analytical role and cross-layer boundary

The outputs of the semantic interpretation framework define the state variables used by later components of the Design Cycle. The Internet Archive layer produces $html_state_i$ and related archival interaction variables, which support the archive-side evidence-state model and transition rules in Section 3.5.4. The YouTube platform layer produces $avail_v^{\text{endpoint}}$ and $avail_v^{\text{html}}$, which support the empirical analysis of platform-facing availability and the **RQ2** regression design.

The two layers remain analytically separate. The Internet Archive layer provides longitudinal archival evidence, while the YouTube platform layer provides a contemporary platform-facing observation. Their outputs are compared in Section 3.5.5, but they are not reconciled into a single final availability state. This boundary preserves the methodological distinction between historical archival evidence and live platform-facing evidence and prevents the analysis from attributing platform intent, moderation decisions, or causal explanations to externally observable signals.

3.5.4 Component M2: Archival Availability Trajectory Reconstruction

Building on the Internet Archive layer specified in Section 3.5.3, this section specifies an algorithm for reconstructing *observed availability trajectories* of archived YouTube videos. The algorithm operates over time-indexed sequences of archival snapshots and produces (i) snapshot-level classifier labels, archive-side evidence states where supported by the observations, and (ii) the event times and indicators required for time-to-event descriptive analysis. Observations that do not provide sufficient evidence for one of the seven interpretable archive-side evidence states are retained as unknown archive evidence rather than being forced into a substantive state as reported in Chapters 4 and 5.

The classifier is developed as a methodological artefact that formalises externally observable archival signals into a reproducible temporal representation. It does not assert causal explanations, platform intent, or definitive states of content existence. Instead, it provides a structured procedure for translating heterogeneous archival evidence into analytically comparable trajectories under conditions of platform opacity.

The detailed derivation and literature positioning of the *7+1 archive-side evidence-state model* are provided in Appendix D.

Inputs and Data Model

For each YouTube video identifier v, the algorithm assumes access to the following archival evidence:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

- A CDX-derived snapshot list

$$S(v) = \{s_i = (t_i,\ status_i,\ mime_i)\}_{i=1}^{n},$$

where t_i denotes the capture timestamp, $status_i$ the HTTP status code recorded by the Internet Archive crawler at crawl time (e.g. 200, 302), and $mime_i$ the recorded MIME type (e.g. `text/html`, `application/binary`). Snapshots are ordered such that $t_1 < t_2 < \cdots < t_n$.

- For snapshots with $mime_i =$ `text/html` and replayable payloads, the archived HTML snapshot H_i and an associated semantic annotation

$$html_state_i \in \{available,\ restricted,\ unavailable,\ eroded\},$$

derived from the interpretation rules applied to archived HTML snapshots. These annotations are derived from observable interface-level signals and metadata properties preserved in the archived representation.

- The timestamp of the last replayable archived HTML snapshot, denoted $t_{\text{last_html}}$, which is used to define analytic event times and to determine right-censoring.

The classifier follows a *two-layered design*: (i) a CDX-only stage, applicable to the full corpus and providing coarse temporal structure; and (ii) an HTML-enhanced stage, which refines evidence-state transitions where replayable archived HTML snapshots are available. This layered design prevents the system from producing interpretations that are not supported by the available observations. HTML-based refinement is applied only when such snapshots exist; Where replayable archived HTML is unavailable, CDX interaction signals may support redirect-only, archival-gap, terminal-absence, or residual unknown archive evidence state. This design reflects the empirical distribution of archival evidence, in which CDX metadata scales corpus-wide, while archived HTML snapshots provide higher-resolution information for a subset of videos.

The 7+1 Archive-Side Evidence-State Model

The *7+1 archive-side evidence-state model* (ASEM_{7+1}) distinguishes seven interpretable archive-side evidence states and one residual unknown archive evidence state within the archival record of a video v. The seven interpretable states are:

State 1. Available (q_{avail}): At least one replayable archived HTML snapshot with $status_i = 200$, $mime_i =$ `text/html`, and $html_state_i = available$, indicating that the video page and embedded content were accessible in the archived representation.

State 2. Restricted (q_{restr}): Replayable archived HTML snapshots with $status_i = 200$, $mime_i =$ `text/html`, and $html_state_i = restricted$, corresponding to interface-level access limitations such as age-gating, region restrictions, or authentication requirements.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

State 3. Eroded (q_{erode}): Replayable archived HTML snapshots with $status_i = 200$, $mime_i = \texttt{text/html}$, and $html_state_i = eroded$, denoting observable loss or degradation of contextual metadata, such as title, description, or channel information, relative to earlier captures.

State 4. Unavailable (q_{unavail}): Replayable archived HTML snapshots with $status_i = 200$, $mime_i = \texttt{text/html}$, and $html_state_i = unavailable$, indicating that the archived page displays a platform-level unavailability notice.

State 5. Redirect-only (q_{redir}): Sequences of at least two consecutive captures with redirect status and no replayable archived HTML payloads, indicating repeated archive-level redirection responses without replayable page content.

State 6. Archival gap (q_{gap}): Periods exceeding a long-gap threshold Δt_{gap} without recorded captures, despite prior or subsequent archival evidence. This evidence state is distinct from the one-day CDX temporal gap count used for structural fragmentation.

State 7. Terminal absence (q_{disp}): A terminal replayability evidence state in which a usable replayable archived HTML snapshot exists, later archive-side observations are present, and no later replayable archived HTML snapshot appears.

In addition to these seven interpretable states, the model retains one residual unknown archive evidence state:

Residual state: Unknown archive evidence (q_{unk})
CDX or processed snapshot evidence exists, but the available evidence is insufficient to assign the observation to one of the seven interpretable archive-side evidence states.

The residual state is part of the evidence-state model, but it is not one of the seven interpretable states. It records the existence of archival evidence without assigning a substantive availability interpretation.

Snapshot-level labels are single-valued: each processed snapshot receives one implemented evidence label. Video-level evidence membership is non-exclusive, because a video may exhibit evidence for more than one archive-side evidence state across its archival record. Archival gap and terminal absence are derived from trajectory structure rather than assigned as ordinary replayable HTML snapshot labels. Missing captures are not imputed.

Finite-State Machine Formulation

Formally, the classifier is represented as a finite-state machine

$$\mathcal{M} = (Q, q_0, \Sigma, \delta),$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

where $Q = \mathcal{Q}_{\text{archive}}$ denotes the evidence-state space of the *7+1 archive-side evidence-state model*, q_0 the initial state, Σ the alphabet of implemented snapshot-level classifier labels, and δ the transition function specified by the rules below. The label `unknown` is retained as a residual unknown archive evidence state. The finite-state formulation is employed as a formal device to ensure transparency and reproducibility and does not imply an ontological model of platform behaviour.

Transition Rules

Let $p(t)$ denote the archive-side evidence state assigned to video v at time t, where assignment is supported by the available observations. Transition rules are defined as follows:

- **Initialisation**: The earliest replayable archived HTML snapshot classified as *available* defines $t_{\text{avail_start}}$. If no such snapshot exists, $t_{\text{avail_start}}$ is undefined. In such cases, the earliest restricted or eroded snapshot provides initial evidence-state information but does not constitute an available onset.
- **Restriction onset**:
$$t_{\text{restr_start}} = \min\{t_i \mid \mathit{html_state}_i = \mathit{restricted}\}.$$
- **Erosion onset**:
$$t_{\text{erode_start}} = \min\{t_j \mid M(H_j) \prec M(H_k),\ t_j > t_k\},$$
where $M(\cdot)$ denotes a metadata vector and $\prec$ denotes component-wise loss.
- **Unavailability onset**:
$$t_{\text{unavail_start}} = \min\{t_i \mid \mathit{html_state}_i = \mathit{unavailable}\}.$$
- **Redirect-only segments**: Sequences of at least k captures (typically $k \geq 2$) with $\mathit{status}_i = 302$ and no replayable archived HTML payloads are labelled P_{redir}.
- **Archival gaps**:
$$\text{If } t_{i+1} - t_i > \Delta t_{\text{gap}},\ \text{label the interval as } P_{\text{gap}}.$$
This archive-side long-gap threshold is separate from the implemented one-day CDX temporal-gap metric used for structural fragmentation.
- **Analytic disappearance time**:
$$t_{\text{disp}} = \begin{cases} t_{\text{unavail_start}}, & \text{if HTML unavailability is observed,} \\ t_{\text{last_html}}, & \text{otherwise.} \end{cases}$$

The disappearance time functions as an event marker for longitudinal analysis and does not assert a factual moment of deletion or platform action.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Algorithm Outline

Algorithm 1 summarises the classification procedure.

Algorithm 1 Observed Archive-Side Evidence-State Classifier for a Single Video

Require: Sorted snapshots $S(v)$ with semantic labels $html_state_i$ where available

Initialise evidence-state timeline $p(t)$ as undefined
Derive baseline metadata $M(H_{\text{base}})$ from earliest available archived HTML snapshot, if present
for each snapshot $s_i = (t_i, status_i, mime_i)$ in $S(v)$ **do**
    if $mime_i =$ `text/html` and $status_i = 200$ **then**
        classify $html_state_i$
        if $html_state_i = available$ and no previous $t_{\text{avail_start}}$ **then**
            $t_{\text{avail_start}} \leftarrow t_i$
        end if
        if $html_state_i = restricted$ and no previous $t_{\text{restr_start}}$ **then**
            $t_{\text{restr_start}} \leftarrow t_i$
        end if
        if $html_state_i = eroded$ and no previous $t_{\text{erode_start}}$ **then**
            $t_{\text{erode_start}} \leftarrow t_i$
        end if
        if $html_state_i = unavailable$ and no previous $t_{\text{unavail_start}}$ **then**
            $t_{\text{unavail_start}} \leftarrow t_i$
        end if
    else if $status_i = 302$ **then**
        mark snapshot s_i as candidate for P_{redir}
    end if
end for
detect redirect-only segments, long archival gaps, terminal replayability absence, and unknown archive evidence in $S(v)$
derive analytic disappearance time t_{disp}
construct evidence-state timeline $p(t)$ where supported by the observations
return evidence-state timeline, residual evidence labels, and event times

Outputs and Derived Analytical Variables

For each video v, the classifier yields:

- Event times:

$$t_{\text{avail_start}}, \quad t_{\text{restr_start}}, \quad t_{\text{erode_start}}, \quad t_{\text{unavail_start}}, \quad t_{\text{disp}}.$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

- A censoring indicator:

$$\delta_v = \begin{cases} 1 & \text{if HTML unavailability is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

- Evidence-state durations or membership indicators $d_v^{(q)}$ for each state $P \in Q$ where sufficient evidence supports assignment.

Formally, the classifier produces the tuple:

$$\mathcal{C}(v) = \big(p(t),\ t_{\text{avail_start}},\ t_{\text{restr_start}},\ t_{\text{erode_start}},\ t_{\text{unavail_start}},\ t_{\text{disp}},\ \delta_v,\ \{d_v^{(q)}\}_{q \in Q}\big),$$

which constitutes the observed evidence descriptor for video v and is the basis for the archive-side analytical variables reported in Chapters 4 and 5. Where no substantive state rule is satisfied, residual unknown evidence is retained rather than forced into one of the seven interpretable states.

For videos without replayable archived HTML snapshots, HTML-based event times such as $t_{\text{avail_start}}$, $t_{\text{unavail_start}}$, and t_{disp} are undefined and are recorded as missing. For these videos, only CDX-based structural indicators such as archival presence, archival lifespan, and CDX temporal gap count are populated.

From $\mathcal{C}(v)$ and $S(v)$, the following scalar variables are derived for use in the analytical feature dataset.

Time-to-event variables. The event-time variables defined below are computed only for videos with a defined $t_{\text{avail_start}}$. The primary outcome variable is the *time to unavailability*:

$$T_v = \begin{cases} t_{\text{unavail_start}} - t_{\text{avail_start}} & \text{if } \delta_v = 1, \\ t_{\text{last_html}} - t_{\text{avail_start}} & \text{if } \delta_v = 0. \end{cases}$$

This formulation accounts for right-censoring when a video has not yet become unavailable at the end of the observation window. When the corresponding evidence-state onsets are observed, additional event-time variables are defined:

$$\begin{aligned} T_v^{(\text{restr})} &= t_{\text{restr_start}} - t_{\text{avail_start}}, \\ T_v^{(\text{erode})} &= t_{\text{erode_start}} - t_{\text{avail_start}}, \\ T_v^{(\text{disp})} &= t_{\text{disp}} - t_{\text{avail_start}}. \end{aligned}$$

These variables are specified as part of the general framework. In the Syria corpus, $T_v^{(\text{restr})}$ is populated for five videos only and $T_v^{(\text{erode})}$ is unpopulated, as documented in Chapter 4 (Section 4.6).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Archival coverage and gap indicators. The following indicators describe structural properties of the archival record of video v and are used as descriptive variables in the RQ1 analytical population, as reported in Chapter 4 (Sections 4.3 and 4.6):

- *Archival presence*: whether any CDX snapshot exists for v ($|S(v)| > 0$), derived from $S(v)$.
- *Archival lifespan*: the elapsed time between the earliest and latest CDX snapshot in $S(v)$, equal to $t_n - t_1$, derived from $S(v)$.
- *CDX temporal gap count*: the number of intervals greater than one day between consecutive CDX snapshots in $S(v)$. This structural fragmentation metric is separate from the Archival gap evidence state q_{gap}, which uses a longer archive-side long-gap threshold.

These indicators are not interpreted as evidence of platform-side behaviour.

3.5.5 Component M3: Cross-Source Observability Boundary Assessment

Purpose and Scope

Both observation layers specified in Section 3.5.3 are structurally independent and produce evidence that cannot be directly merged or reconciled. This section specifies a procedure that makes the relationship between their outputs explicit and analytically tractable by documenting four conditions of joint observability: (i) where Wayback CDX records are accompanied by archived HTML snapshots at the same archival timestamp; (ii) whether those archived HTML snapshots expose a recoverable player response; (iii) whether archived HTML snapshots with recoverable semantic signals exist prior to the live check; and (iv) how the endpoint-derived availability states from the YouTube platform layer and the playability signals from the Internet Archive layer are jointly distributed across the corpus.

Non-alignment between layers is treated as an empirical property of the measurement infrastructure rather than as evidence of platform-side behaviour, content removal, or moderation intent. All results of this procedure are reported in Section 4.7.

Inputs

The procedure operates on three inputs per video identifier v, each formally defined in Section 3.5.3:

- The Wayback CDX snapshot metadata list from the Internet Archive layer:

$$S_{\text{CDX}}(v) = \{\, s_i = (t_i,\ status_i,\ mime_i) \,\}_{i=1}^{n}.$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

- The archived HTML snapshot set from the Internet Archive layer, where each snapshot carries two signal fields:

$$S_{\text{HTML}}(v) = \{\, h_j = (t_j,\ \textit{html_state}_j,\ \textit{playability_state}_j,\ \textit{has_signal}_j) \,\}_{j=1}^{m},$$

  where:

  - t_j is the archival capture timestamp,
  - $\textit{html_state}_j \in \{\textit{available}, \textit{restricted}, \textit{unavailable}, \textit{eroded}\}$ is the evidence label assigned by the semantic interpretation layer (Section 3.5.3), used in the availability trajectory reconstruction (Section 3.5.4),
  - $\textit{playability_state}_j$ is the raw playability signals extracted from the player response embedded in the archived HTML snapshot — values include `OK`, `LOGIN_REQUIRED`, `ERROR`, and *no player response* — and is used in the cross-source joint distribution (Stage 3 below),
  - $\textit{has_signal}_j \in \{0, 1\}$ indicates whether a recoverable player response was present in the archived HTML snapshot, with $\textit{has_signal}_j = 1$ corresponding to snapshots classified as *Captured YouTube HTML* with a detectable player response, and $\textit{has_signal}_j = 0$ corresponding to snapshots with no recoverable player response — including Wayback rate-limit notices, captured non-YouTube HTML, captured site error pages, and Wayback robots exclusion pages.

- The YouTube platform layer output $Y(v) = (t_{\text{check}},\ \textit{avail}_v^{\text{endpoint}},\ \textit{avail}_v^{\text{html}})$ as defined in Section 3.5.3. This procedure uses only the endpoint-derived availability field $\textit{avail}_v^{\text{endpoint}}$ and the request timestamp t_{check}. The HTML-derived field $\textit{avail}_v^{\text{html}}$ is not used here; its analytical role is specified in Section 3.5.3.

Procedure

The assessment proceeds in four stages, each producing a distinct observability indicator.

Stage 1 — CDX–archived HTML timestamp overlap. For each Wayback CDX snapshot record $s_i \in S_{\text{CDX}}(v)$, the procedure checks whether an archived HTML snapshot retrieved via the Wayback replay layer exists at the same archival timestamp:

$$\textit{html_match}(s_i) = \begin{cases} 1 & \text{if } \exists\, h_j \in S_{\text{HTML}}(v) \text{ such that } t_j = t_i, \\ 0 & \text{otherwise.} \end{cases}$$

A CDX record with $\textit{html_match}(s_i) = 0$ indicates that the Internet Archive indexed a capture attempt at timestamp t_i but no corresponding archived HTML snapshot is available for replay at that timestamp. Such records are treated as index-level archival evidence only and do not support semantic availability classification at that timestamp.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Stage 2 — Semantic yield (archived html) conditional on CDX–HTML overlap. For each archived HTML snapshot h_j with $html_match(s_i) = 1$ at the corresponding CDX timestamp, the procedure determines whether the snapshot exposes a recoverable player response:

$$has_signal_j = \begin{cases} 1 & \text{if a player response is extractable from } h_j, \\ 0 & \text{otherwise.} \end{cases}$$

A matched snapshot with $has_signal_j = 0$ indicates that the archived HTML snapshot exists at the timestamp but its content does not expose YouTube-specific semantic evidence — for example, because the Wayback replay layer returned a rate-limit notice, a robots exclusion page, a captured site error page, or a non-YouTube page instead of the original platform content. The existence of a replayable archived HTML snapshot therefore does not guarantee semantic yield (archived html).

Stage 3 — Archived HTML semantic signal availability prior to live check. For each video v, the procedure determines whether any archived HTML snapshot with a recoverable player response exists at or before the live check timestamp t_{check}:

$$has_prior_semantic(v) = \begin{cases} 1 & \text{if } \exists\, h_j \in S_{\text{HTML}}(v) \text{ such that } t_j \le t_{\text{check}} \text{ and } has_signal_j = 1, \\ 0 & \text{otherwise.} \end{cases}$$

When $has_prior_semantic(v) = 0$, no archived HTML evidence with a recoverable semantic signal is available from the Internet Archive layer to contextualise the live platform check outcome. This constitutes a structural observability boundary: the endpoint-derived availability state cannot be compared against any prior archival semantic evidence for this video.

Stage 4 — Joint distribution of archived HTML playability states and endpoint-derived availability states. For each video v with $has_prior_semantic(v) = 1$, the procedure selects the latest archived HTML snapshot with a recoverable player response prior to the live check:

$$h^*(v) = \arg \max_{\substack{h_j \in S_{\text{HTML}}(v) \\ t_j \le t_{\text{check}} \\ has_signal_j = 1}} t_j.$$

The cross-source observation for video v is the pair:

$$(playability_state_{h^*(v)},\ avail_v^{\text{endpoint}}),$$

where $playability_state_{h^*(v)}$ is the playability signals extracted from the player response of the latest archived HTML snapshot prior to the live check, and $avail_v^{\text{endpoint}}$ is the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

endpoint-derived availability state from the YouTube platform layer at the time of measurement.

These pairs are aggregated into a joint distribution table across the corpus. For videos where $has_prior_semantic(v) = 0$, no pair is formed. These videos are recorded separately as cases where cross-source comparison is structurally impossible due to the absence of prior archival semantic evidence.

Outputs

The procedure produces four outputs per corpus execution:

(i) A *CDX–archived HTML overlap record*: for each Wayback CDX snapshot record, the boolean indicator $html_match(s_i)$ recording whether a Wayback-replayed archived HTML snapshot exists at the same archival timestamp.

(ii) A *semantic yield (archived html) record*: for each archived HTML snapshot with $html_match(s_i) = 1$, the boolean indicator has_signal_j recording whether a recoverable player response is present, together with the distribution of response shape categories across matched snapshots.

(iii) A *structural observability boundary indicator*: $has_prior_semantic(v) \in \{0, 1\}$ per video, recording whether any archived HTML snapshot with a recoverable player response exists in the Internet Archive layer prior to the live YouTube platform check.

(iv) A *joint state distribution table*: for each video with $has_prior_semantic(v) = 1$, the pair of the playability signals from the latest archived HTML snapshot prior to the live check and the endpoint-derived availability state from the YouTube platform layer.

Analytical Treatment of Non-Alignment

Divergence between archived HTML playability states and endpoint-derived availability states is not treated as measurement error or data quality failure. It is interpreted as an empirical property of the observation infrastructure arising from four structurally distinct sources:

- *Temporal distance between layers*: the latest archived HTML snapshot may precede the live check by months or years, during which the platform-facing availability state may have changed independently of any change observable in the Internet Archive layer.
- *Wayback replay-layer constraints*: as specified in Section 3.5.3, archived HTML snapshots may consist of replay-layer-generated interstitial pages, robots exclusion

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

responses, or captured site error pages that carry no YouTube-specific semantic information, producing $has_signal_j = 0$ even when a CDX record exists.

- *CDX indexing without replay availability*: a Wayback CDX snapshot metadata record documents that a capture attempt occurred but does not guarantee that the corresponding archived payload is available for replay at the time of retrieval.

- *Non-target platform responses during live checks*: as specified in Section 3.5.3, when the YouTube platform returns a non-target HTML response — such as a rate-limiting interstitial — the HTML-derived signal yields *indeterminate* and the endpoint-derived signal yields *unknown*. Both signals converge on unobservability for the same request, and such videos cannot contribute to the joint distribution in Stage 4.

All four sources of non-alignment are properties of the Internet Archive and YouTube platform observation infrastructures, not of platform-side behaviour. Non-alignment is therefore retained as an explicit analytical outcome and is reported as a corpus-level observability limit in Chapter 4 (Section 4.7). It is not interpreted as evidence of content removal, moderation action, or platform-side change between the archival capture and the live check.

3.5.6 Component M4: Human Taxonomy Features Construct

In addition to availability trajectories derived from archival evidence, the Design Cycle constructs content-related categorical covariates from human-assigned taxonomy labels. These variables serve as fixed explanatory variables in regression analyses addressing **RQ2**. Their construction is completed before statistical analysis and forms part of the frozen preprocessing stage of the observational measurement system. The construct does not infer content meaning, political stance, or narrative intent at the level of individual videos. It operates on existing human-provided labels and transforms them through a fixed sequence of explicit preprocessing steps: field-based axis separation, manual semantic-core mapping, corpus-level diagnostic evaluation, admissibility filtering, and per-video feature assignment. Each step is documented, deterministic, and reproducible.

The full taxonomy mapping used to derive semantic-core and semantic-top features is documented in Appendix F.

Human Label Inputs

Each video identifier may be associated with zero or more human-assigned taxonomy labels originating from the curated corpus. These labels are treated as exogenous inputs. At ingestion time, they are preserved in their original form and are not discarded, reweighted, or replaced. Videos without any human-assigned labels are retained explicitly as unlabelled cases and are not excluded by default.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The construct uses three taxonomy fields for content-feature generation: `main_classification`, `group`, and `subcategory`. These fields provide the input label space for the three feature axes used in **RQ2**. Other metadata fields may remain available in the broader pipeline, but they do not enter this construct.

Axis-Based Decomposition

To create a structured and auditable feature space, labels are separated into three fixed axes defined by their source field:

- **Axis A (Main):** labels from `main_classification`
- **Axis B (Group):** labels from `group`
- **Axis C (Subcategory):** labels from `subcategory`

This step is purely structural. It does not change the labels. It only assigns each label to the axis defined by its original field. Labels are not moved across axes, and no label from one field is reassigned to another axis. A video may have labels on any subset of the three axes, including none. The axes are therefore treated as separate structural dimensions rather than as a single combined taxonomy.

Semantic Core Normalisation

Raw human-assigned labels exhibit lexical variation, synonymy, partial overlap, and high sparsity across the corpus. A direct use of raw labels would therefore produce an unstable and fragmented predictor space. To address this problem, a manual *semantic core normalisation* step is applied independently within each axis.

In this step, each raw taxonomy label is mapped to a manually defined *semantic core*. This mapping is many-to-one: multiple raw labels may map to the same semantic core. The purpose of this step is to reduce variation in surface terminology and to create a stable canonical label space for later feature construction. For example, raw labels referring to the aftermath of shelling and destruction, bombing or targeting events, and instances of gunfire are mapped to the shared semantic core *Armed violence and combat*.

This step contains a substantial amount of manual work. The semantic cores are not discovered automatically by a model. They are defined in advance through documented mapping tables. Constructing these tables requires explicit analytic judgement about which raw labels should be treated as substantively equivalent for the purposes of feature construction. The work performed here is therefore not a neutral technical clean-up step. It is a manual consolidation of the raw taxonomy into a smaller set of canonical categories that remain grounded in the curated label system.

The mapping process does not rely on NLP, clustering, or statistical inference. It does not create labels from text and it does not infer video-level categories from NLP or

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

statistical models. Instead, it consolidates existing human-assigned labels into canonical semantic cores. Raw labels are preserved in parallel for traceability, while the mapped semantic cores are used for downstream processing.

If a video contains at least one raw label that maps to a semantic core on a given axis, that axis is considered populated at the semantic-core level. Where no mapped semantic core is available on an axis, the axis remains explicitly unassigned. Missingness is therefore preserved rather than filled through imputation.

Label–Mapping Consistency and Retention

After semantic-core mapping, the dataset is checked for structural consistency. This check does not remove unlabelled videos. Videos without any human-assigned labels remain included as unlabelled observations. Videos with human-assigned labels also remain in the dataset even if they do not yield an admissible final feature on one or more axes. In such cases, the corresponding axis value remains missing.

This point is important for transparency. The construct does not assume that every labelled video must produce a final usable feature on every populated axis. A video may retain its raw labels and mapped semantic cores and still fail to receive a final semantic-top assignment on a given axis because the mapped category does not satisfy the later admissibility criteria. These cases are not hidden or silently repaired. They remain in the dataset and are represented explicitly through missing values in the relevant feature columns.

Corpus-Level Diagnostic Evaluation

Once semantic cores have been defined, they are evaluated at the corpus level to determine whether they are sufficiently frequent and sufficiently stable for use as regression predictors. This step uses descriptive diagnostics rather than video-level semantic inference.

The diagnostics are computed independently within each axis and include:

- label frequency,
- topic–label co-occurrence distributions,
- topic agreement scores,
- label dispersion across topics,
- entropy-based divergence measures, and
- coverage and missingness patterns.

At this stage, NLP is used only as an auxiliary diagnostic tool. Text-based topic modelling is applied to the corpus in order to evaluate how mapped labels are distributed across

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

broader text clusters. These diagnostics are treated as properties of labels and their corpus-level behaviour, not as properties of individual videos. They are not used to define the semantic cores themselves, and they are not used to assign meaning to individual observations.

This diagnostic stage includes an explicit modelling rule. Quantitative outputs are translated into stability classes such as *stable*, *mixed*, *fragmented*, and *sparse*. The rule is deterministic once specified, but the choice to use these classes, and to use them for admissibility filtering, is a modelling decision.

Admissibility Filtering

Not every semantic core is retained as a regression predictor. After the diagnostic step, each semantic core is evaluated against fixed admissibility criteria. In the current implementation, a semantic core is retained only if it meets both of the following conditions:

- it satisfies a minimum frequency threshold in the corpus, and
- its stability class is admissible for modelling.

In the frozen implementation used for this thesis, the minimum frequency threshold is fixed at `50`, and the admissible stability classes are `stable` and `mixed`. Semantic cores that do not satisfy both conditions are not used as final regression predictors.

This step is not hidden preprocessing. It is a direct modelling decision that reduces the final predictor space to categories with sufficient empirical support and acceptable corpus-level stability. The purpose of this reduction is to avoid extremely sparse or highly fragmented predictors that would weaken the interpretability and stability of the regression models.

Semantic Top Identifiers

For each admissible semantic core, the construct assigns an axis-specific *semantic top identifier*. In the final implementation, semantic tops do not represent an additional higher-order grouping layer above multiple semantic cores. Instead, they function as fixed identifiers for the admissible semantic cores that remain after filtering.

Each semantic top is therefore defined by:

- a unique identifier,
- axis membership, and
- a frozen admissibility status determined by the diagnostic and frequency constraints described above.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Once these identifiers are constructed, they remain fixed throughout the analysis. No subsequent step redefines them, merges them, or alters their admissibility status. This freezing step ensures that the feature space remains constant across all downstream analyses.

Per-Video Assignment

For each video identifier, semantic-top features are assigned independently for each axis. The assignment rule is direct:

- If a video contains an admissible semantic core on an axis, the corresponding semantic-top identifier is assigned.
- If no admissible semantic core is present on that axis, the feature value is recorded as missing.

This results in up to three categorical features per video:

- `axis_A_main__semantic_top_id`
- `axis_B_group__semantic_top_id`
- `axis_C_subcat__semantic_top_id`

The observation unit remains the individual video. No row duplication, no cross-axis pooling, and no imputation are performed. A video may receive a feature on all three axes, on some axes only, or on none. In all cases, the resulting feature representation reflects the actual output of the preceding manual and diagnostic steps rather than a completed or repaired version of the data.

Role of NLP-Derived Diagnostics

Text-based clustering and NLP methods are not used to define the semantic meaning of individual videos and are not included as regression covariates. Their function is limited to corpus-level diagnostic support during admissibility assessment. More specifically, NLP-derived topic structures are used to evaluate the internal coherence, dispersion, and coverage of mapped semantic cores. They help identify whether a manually defined category behaves in a sufficiently stable way across the corpus to justify its use as a predictor.

Once admissibility decisions and semantic-top identifiers are fixed, raw NLP outputs are not used in the final regression models. Topic assignments, cluster memberships, and similarity scores are discarded as predictors. This design preserves the primacy of human-assigned labels while still making use of corpus-level diagnostics to evaluate whether the manual taxonomy can support stable quantitative analysis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Analytical Role in RQ2

The final output of this construct is a set of axis-specific categorical features derived from human-assigned taxonomy labels through explicit manual mapping, diagnostic evaluation, and admissibility filtering. These features are used as explanatory variables in the regression models for **RQ2**. They represent content-related metadata dimensions in a form that is auditable, frozen, and aligned with the observable constraints of the measurement system.

This explicit construction process is important methodologically. It makes the manual work behind the taxonomy visible, shows where analytical judgement enters the pipeline, and avoids presenting the final variables as if they emerged automatically from the data. At the same time, it avoids overclaiming. The resulting variables do not capture the full meaning of the videos and do not exhaust the interpretive richness of the curated labels. They provide a structured and reproducible representation of selected content dimensions for statistical analysis under clearly defined constraints.

3.5.7 Component M5: Archival Completeness Construct

Building on the archival-first logic defined above and the conceptual foundations outlined in Section 2.3, this subsection operationalises *completeness* as a structured, artefact-based construct that captures the degree to which a video is archivally represented across multiple preservation layers. The construct is derived from externally observable archival evidence and is designed to transform observable preservation traces into an analytically usable measure. It follows measurement-theory principles and Information Systems construct-validity guidance for defining reproducible indicators from observable system artefacts [Ste46, SBG04, HC10, RA18, KSBVdS19].

Completeness is derived exclusively from *externally observable archival evidence* produced by the measurement system. No inference about user intent, platform motivation, or content meaning is introduced at this stage. The operationalisation follows explicitly defined calculation rules and can be repeated using the same inputs to obtain identical results. The resulting measure reflects the observable state of archival capture rather than interpretive judgement.

Completeness does not measure importance, relevance, or value of content. It formalises the limits of what can be known and reconstructed from archival evidence, thereby constraining downstream analytical claims to what the preservation infrastructure actually supports.

Design constraints

The completeness construct is governed by five design constraints. First, it operates only on externally observable archival signals. Second, it represents archival preservation as a graded condition rather than as a binary available/unavailable state. Third, it preserves structural traces rather than attempting to recover missing content or infer platform-side

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

causes. Fourth, it is computed independently of content categories, political framing, or preservation priority. Fifth, it is reproducible: identical inputs and fixed calculation rules produce identical outputs.

These constraints ensure that the construct remains an observational artefact. It records what archival evidence exists, what form that evidence takes, and where observability fails, without converting incompleteness into claims about content value, deletion cause, or platform intent.

Archival layers and modal preservation states

At its core, completeness is determined by the *modal availability of archival layers* associated with each video identifier. Three archival layers are considered: Wayback CDX records, replayable HTML snapshots, and WARC payload captures. Their presence or absence defines a hierarchy of preservation states ordered by informational richness and reconstructive capacity. This hierarchy constitutes the primary qualitative dimension of the completeness construct.

Table 3.2: Operationalisation of completeness from observable archival evidence to preservation level and ordinal score.

Archival layers	Observed condition	State	Score	Interpretation
WARC + HTML + CDX	All technical and contextual traces captured; artefact fully reconstructable	**Full**	1.0	Complete archival persistence; content and provenance jointly recoverable
HTML + CDX	Public-facing representation preserved; archival payload missing	**Partial**	0.7	Substantial but degraded preservation; representation retained without full fidelity
CDX only	Index-level reference without replayable content	**Trace**	0.3	Residual archival footprint; evidentiary trace without reconstructable artefact
No archival record	No archival artefact observed	**Absent**	0.0	No detectable archival persistence; artefact unobservable

The numeric values associated with each preservation state serve as *ordinal encodings* that preserve rank order without implying equal intervals or linear distances between categories. They enable systematic comparison and aggregation while remaining consistent with the qualitative nature of archival preservation states. Accordingly, completeness is interpreted as a *graded preservation status* rather than as a precise quantitative magnitude.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Secondary quantitative modifiers

In addition to the modal preservation state, two secondary quantitative characteristics are recorded to refine the empirical description of archival completeness: (i) the number of archived snapshots associated with a video identifier and (ii) the temporal span in days between the first and last observed snapshot. These characteristics are computed only when replayable archival content is present and are explicitly gated by content availability.

These secondary components do not define additional preservation categories. Instead, they function as *within-state modifiers* that capture variation in archival density and temporal coverage among videos sharing the same modal preservation state. Their contribution is bounded and strictly constrained by the modal preservation state, ensuring that the absence of core archival layers cannot be compensated for by temporal or volumetric factors alone.

Aggregation rule

To obtain a scalar completeness indicator suitable for regression modelling, the modal preservation state and the two secondary modifiers are combined through a fixed weighted aggregation:

$$S_v = 0.60 \cdot M_v + 0.25 \cdot C_v + 0.15 \cdot T_v,$$

where:

- M_v denotes the ordinal score of the modal archival state,
- $C_v = \min\left(\frac{\text{snapshot_count}_v}{10}, 1\right)$ denotes the normalised snapshot-count component,
- $T_v = \min\left(\frac{\text{span_days}_v}{180}, 1\right)$ denotes the normalised temporal-span component.

Both secondary components are computed only when replayable HTML or WARC content is present; otherwise, $C_v = 0$ and $T_v = 0$. The weights are fixed and applied uniformly across all video identifiers. The final value is bounded by the modal preservation score such that:

$$S_v = \min(0.60 \cdot M_v + 0.25 \cdot C_v + 0.15 \cdot T_v,\ M_v)\,.$$

This bounding rule ensures that variation in snapshot count or temporal span cannot elevate a video beyond its modal preservation state. The resulting scalar indicator preserves the ordinal hierarchy of archival layers while enabling structured comparison within preservation categories.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Validation and analytical suitability

This operationalisation satisfies established criteria for construct validity in Information Systems research. The construct exhibits clear correspondence between empirical indicators and conceptual meaning, transparent derivation rules, and reproducible computation [SBG04]. By separating modal archival presence from secondary quantitative modifiers, the design avoids conflating categorical preservation states with temporal persistence or snapshot frequency.

The resulting completeness measure supports subsequent empirical analyses that examine distributions, correlations, and trajectories of archival persistence, while remaining aligned with the stated limits on what the analysis claims and the ethical commitments outlined in Chapter 2. Completeness therefore functions as a system-level indicator of digital-memory preservation grounded in observable archival artefacts rather than inferred content significance.

The derivation, literature positioning, and boundary conditions of the completeness artefact are documented in Appendix E.

Analytical use and non-use

The completeness construct is used as an interpretive qualifier for downstream analysis. It conditions how availability-related observations and statistical associations are interpreted by making the degree and form of archival evidence explicit.

The construct may support analytical stratification and robustness checks. Completeness levels can be used to compare subsets of videos with different degrees of archival observability, while the scalar score can be used to assess whether observed statistical patterns remain stable under different archival coverage conditions. In these uses, completeness does not explain availability outcomes; it defines the observational conditions under which such outcomes can be interpreted.

The construct is not used as a proxy for content persistence, content disappearance, platform moderation, content importance, or preservation priority. A low completeness score does not indicate that a video was removed, suppressed, or less important. A higher completeness score does not imply uninterrupted platform availability or greater content value. Completeness is also not used as a hard exclusion criterion. Videos with low or absent archival completeness remain part of the analytical population where the relevant research question permits their inclusion.

Design relevance and boundary conditions

Beyond its role as a measurement variable, the completeness construct also functions as a transparency artefact. By representing archival observability as a graded, explicitly specified, and reproducibly calculable condition, it makes visible the preservation limits that shape what can and cannot be reconstructed from archival evidence. This visibility function aligns with the broader orientation of Digital Humanism toward transparency in

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

research and technological design: Principle 3 holds that effective regulations must ensure accountability and transparency of software programs and algorithms [WLA+19]. In this thesis, the completeness construct operationalises this principle at the level of archival evidence rather than platform governance. It transforms the structural conditions of preservation, including archival crawl priorities, replay-layer constraints, and technical indexing decisions, into an explicit and comparable indicator that can be used in analysis without requiring insider access to the systems that produced those conditions.

The construct remains bounded by the evidence available to the observational measurement system. It does not explain why archival layers are present or absent, and it does not attribute responsibility for preservation gaps. Its function is to make those gaps analytically visible, reproducibly measurable, and explicitly available for interpretation in relation to the thesis research questions.

3.5.8 Component M6: Quantitative Observational Analysis

The quantitative observational analysis operationalises the Design Cycle construction described in Section 3.5. It integrates Wayback Machine CDX/WARC traces with YouTube metadata for the sampled set of Syria-related video identifiers (2011–2025), characterises longitudinal patterns of content availability and archival persistence, and examines statistical associations between content-taxonomy attributes and observable availability outcomes. The data work procedure follows CRISP-DM, while reproducibility instrumentation follows PRIMAD [FFJ+16].

Purpose and analytical scope

The purpose of this component is to produce datasets and analytical artefacts that support the empirical analysis of **RQ1** and **RQ2**. For **RQ1**, the component characterises aggregate patterns of content availability, archival presence, temporal coverage, and observed unavailability. For **RQ2**, it tests statistical associations between fixed content-taxonomy attributes and observable availability outcomes. The outputs also provide the empirical basis for the system-level evaluation and design implications addressed under **RQ3**.

The component remains observational. It does not infer platform intent, moderation rationale, or causal mechanisms from the observed data. All quantitative outputs are therefore interpreted as descriptions of externally observable availability and archival evidence rather than as explanations of why specific videos became unavailable or remained available.

Inputs

The quantitative analysis operates on the fixed analytical subset defined in Section 3.3.3. For each video identifier, the following inputs are used:

- the fixed list of sampled YouTube video identifiers;

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- Wayback CDX capture timelines, including first observed capture, recorded MIME types, and recorded HTTP status codes;
- WARC availability and replayable archived HTML evidence, where available;
- YouTube platform-facing availability observations produced by the YouTube platform layer;
- channel and video metadata available in the curated corpus; and
- content-taxonomy features constructed in Section 3.5.6.

Processing procedure

The procedure consists of four stages.

Timeline integration. Video identifiers are joined to their Wayback CDX capture timelines to derive archival coverage, capture continuity, temporal span, and state-transition indicators. This stage produces the temporal variables used to describe observed availability and archival persistence.

Metadata and feature merge. The derived archival variables are merged with platform-facing metadata and the taxonomy-based content features constructed in Section 3.5.6. Categorical variables are normalised according to the frozen preprocessing rules defined for the analytical feature matrices. No post-hoc recoding is introduced after the feature matrices are fixed.

Modelling and descriptive analysis. The resulting datasets are used to characterise time-to-event distributions of observed unavailability and to test statistical associations between observable availability outcomes and content-taxonomy attributes. Descriptive summaries support **RQ1**; regression models support **RQ2**. Seeds, parameters, model specifications, and analytical outputs are recorded for reproducibility.

Logging and manifest generation. All data transformations, parameters, seeds, software versions, and execution environment details are persisted in run manifests. These manifests provide the link between the implemented measurement system, the generated datasets, and the empirical results reported in later chapters.

Methodological justification for live infrastructure interaction

While the technical implementation of the Design Cycle centres on the integration of YouTube metadata and Wayback Machine traces, the decision to interact directly with live preservation infrastructures carries methodological significance. This design enables the empirical observation of meta-preservation: the ways in which archival systems record and expose their own processes of remembering and forgetting. Unlike approaches that

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

rely on static archival corpora such as Common Crawl [Com25] or precompiled Internet Archive index dumps [Int25a], this thesis treats the live responses of the Wayback Machine as data in their own right.

Patterns of rate limiting, request throttling, or temporal gaps in capture activity are interpreted as observable properties of archival operation rather than as explanations of underlying causes. Such dynamics are not directly observable in offline datasets, which present digital memory as a static snapshot rather than as an evolving socio-technical process. Controlled and transparent live access is therefore required to document these operational characteristics within the scope of the observational methodology.

Archival-first data completeness logic

To minimise bias from partial archiving, the measurement system applies an archival-first completeness logic with controlled fallbacks.

1. **Primary path:** parse WARC-based structured metadata and capture events when available.
2. **Fallback path:** if WARC evidence is unavailable or incomplete, parse replayable archived HTML snapshots.
3. **Last-resort path:** if neither WARC nor replayable HTML evidence supports extraction, mark the relevant fields as missing and propagate explicit `source_type` values (`warc`, `html`, `fallback`) together with the scalar `completeness_score`.

This logic is recorded in the run manifest and the database. It enables sensitivity checks on coverage assumptions and preserves explicit missingness rather than silently repairing incomplete archival evidence.

Outputs

The component produces three classes of outputs:

- dataset artefacts linking platform metadata, archival traces, availability variables, completeness indicators, and taxonomy-derived features;
- analytical artefacts, including descriptive time-to-event summaries and regression model outputs; and
- reproducibility artefacts, including run manifests, parameter records, seeds, software-version records, and environment descriptions.

These outputs form the direct empirical basis for the results reported in Chapters 4 and 5.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Checks and analytical boundaries

The component applies rule-based data transformations, schema validation, missingness reports, and unit checks on time fields such as capture dates. Repeated runs with identical inputs, seeds, and parameters are expected to produce identical derived datasets and model outputs.

The analytical boundaries of the component follow from the observational design of the thesis. The analysis can describe distributions, temporal patterns, completeness conditions, and statistical associations in the observed data. It cannot establish causal effects, platform intent, or definitive reasons for content unavailability.

3.6 Design Cycle: Artefact Evaluation

3.6.1 What Is Being Evaluated

The evaluation phase of the Design Cycle assesses the research artefacts produced in Section 3.5 against requirements derived from the problem context established in the Relevance Cycle [HMPR04, Hev07]. The primary artefact under evaluation is the observational measurement system in its role as an *instantiation*: an implemented system that demonstrates the feasibility of making platform-mediated digital memory empirically observable under conditions of platform opacity and archival incompleteness. Three supporting artefacts are evaluated in relation to the instantiation: the *7+1 archive-side evidence-state model* and graded completeness representation as *constructs*; the finite-state machine formulation as a *model*; and the trajectory reconstruction procedure and feature extraction procedure as *methods* [HMPR04, MS95]. The central evaluative question is whether the artefact system, taken together, adequately addresses the four research gaps that motivated its construction (G1–G4, Section 1.3) and satisfies the acceptance criteria defined by the requirements of external, reproducible, longitudinal observation of platform content availability.

3.6.2 Evaluation Methods

Hevner et al. identify five categories of DSR evaluation method: observational, analytical, experimental, testing, and descriptive [HMPR04]. The selection of methods for this thesis is determined by the nature of the artefact, the stage of the research, and the structural constraints of the problem domain. The justification for each decision is as follows.

Analytical evaluation. The artefact has a formal specification that is amenable to static analysis. The finite-state machine $\mathcal{M} = (Q, q_0, \Sigma, \delta)$ can be examined for completeness of evidence-state coverage and consistency of transition rules. The completeness construct has a provable structural property — the bounding rule $S_v \leq M_v$ — that ensures secondary quantitative modifiers cannot elevate a video beyond its modal preservation state. The two-layer CDX/HTML architecture can be assessed for fitness against the empirical distribution of available archival evidence. Analytical evaluation

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

is the primary method for demonstrating that the artefact is internally consistent and structurally sound.

Functional testing. The artefact is an implemented system. Functional testing verifies that the system produces contract-valid and reproducible outputs from specified inputs. Testing is conducted through: (i) an independent re-run on a held-out subset to verify that identical inputs produce identical outputs across execution environments; (ii) schema validation against predefined data contracts; and (iii) PRIMAD manifest checks that record and verify seeds, software versions, and checksums for each run [FFJ+16]. These tests address the reproducibility requirement directly, which is a primary acceptance criterion for a methodological artefact of this kind.

Experimental evaluation. Simulation-style evaluation, in which the system is applied to the full corpus and aggregate outputs are examined for consistency with expected properties, is feasible and is conducted. Controlled experiment against a known ground truth is not applicable: no independently verified ground truth for item-level platform availability over the observation window exists externally. This is not a flaw in the design; it is a boundary condition of the problem domain — the inaccessibility of ground truth is itself a consequence of the platform opacity that the artefact was designed to expose.

Descriptive informed argument. For a methodological instrument operating under conditions where no external benchmark exists, a structured, evidence-grounded argument for utility is both appropriate and necessary [HMPR04]. The descriptive informed argument evaluates artefact adequacy by relating observable artefact properties to the requirements derived from the research gaps G1–G4 and to the design principles established in Section 3.5. In this thesis, the descriptive informed argument is operationalised through structured criteria mapping: evidence from the analytical evaluation, functional testing, simulation-style corpus application, and empirical results is mapped to criteria C1–C5. This mapping makes the adequacy argument explicit, auditable, and traceable. It also provides the evaluation matrix from which system-level design implications and remaining evaluation boundaries are derived. The argument is grounded exclusively in evidence produced by the implemented system and the empirical results of applying it to the corpus.

3.6.3 Evaluation Criteria Applied Through the Structured Mapping

The evaluation criteria define the standards used to judge whether the artefact system satisfies the requirements derived from the Relevance Cycle. They are distinct from the evaluation methods specified above: the methods check, produce, and organise evaluation evidence, while criteria C1–C5 provide the basis for judging the adequacy of the observational measurement system. Each criterion corresponds directly to one or more of the four research gaps that motivated the design.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

C1 — Observability differentiation (addresses G3): the artefact must distinguish qualitatively different states of archival observability without collapsing them into a binary available/unavailable signal.

C2 — Preservation of partial evidence (addresses G1 and G2): the artefact must retain and represent partial archival traces, indeterminate states, and right-censored trajectories as explicit analytical outcomes rather than discarding them.

C3 — Analytical operability without causal attribution (addresses G1): indicators derived from the artefact must be usable in quantitative models without implying platform intent, moderation decisions, or causal explanations.

C4 — Reproducibility and transparency (addresses G4): applying the artefact to identical inputs must produce identical outputs; all transformations must be explicitly recorded and independently verifiable.

C5 — Alignment with problem framing (addresses G1–G4): the artefact must make infrastructural conditions of digital memory observable without embedding normative judgments, content valuations, or causal attributions into the system.

3.6.4 Application in This Thesis

Purpose. The purpose of the evaluation is to determine whether the observational measurement system adequately addresses research gaps G1–G4 and satisfies criteria C1–C5, using the evaluation methods specified above.

Inputs. The evaluation uses the implemented artefact system, dataset and trajectory artefacts, PRIMAD run manifests, and the formal artefact specifications in Section 3.5.

Procedure. The evaluation is conducted in four connected steps. First, analytical evaluation reviews the formal specification of the artefact system, including the finite-state formulation, evidence-state coverage, transition logic, two-layer CDX/HTML architecture, and completeness bounding rule. Second, functional testing verifies implementation behaviour through independent re-run, schema validation, and PRIMAD manifest checks. Third, simulation-style corpus application examines the behaviour of the system on the fixed analytical corpus and checks whether the resulting outputs preserve differentiated observability states, partial traces, right-censoring, missingness, reproducible feature construction, and non-causal interpretability. Fourth, the descriptive informed argument is executed through structured criteria mapping: evidence from the preceding evaluation steps and from the empirical results reported in Chapters 4 and 5 is mapped to criteria C1–C5. The resulting evaluation matrix provides the basis for deriving system-level design implications and for recording remaining evaluation boundaries.

Outputs. The evaluation produces three outputs: (i) an evaluation matrix mapping criteria C1–C5 to evidence from the artefact system, the evaluation activities, and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the empirical results; (ii) system-level design implications derived from the evaluated measurement system and the observed limits of external evidence; and (iii) a discrepancy log recording unresolved boundaries of the evaluation.

Where reported. The execution and results of this evaluation are reported in Section 5.6, after the analytical outputs for RQ1, RQ2, and RQ3 have been presented.

3.7 Reproducibility and Ethical Considerations

Reproducibility Framework. Reproducibility is operationalised through the PRIMAD framework [FFJ+16], which structures the study around six dimensions: *Provenance* (inputs, seeds), *Reference* (versions, checksums), *Instrument* (software, environment), *Method* (procedures, parameters), *Actor* (roles), and *Data* (sources, licences). Each run produces a timestamped manifest together with a pass/fail validation report covering schema consistency and missingness.

Applying this framework to a study of platform-mediated digital memory means operationalising, at the level of the research infrastructure itself, the transparency and accountability requirements that Principle 3 identifies as structural necessities for socially responsible technological systems: software programs and algorithms should ensure prediction accuracy, fairness, accountability, and transparency [WLA+19]. The explicit recording of provenance, instrument versions, methods, and data sources across runs is therefore not merely a technical reproducibility measure. It makes the conditions and consequences of algorithmic processes—in this case, the process of measurement itself—visible, verifiable, and independently auditable. This further instantiates Principle 8's requirement that researchers engage openly with wider society and reflect critically on their approaches.

Ethical Boundary Conditions. Ethical considerations inform several design choices throughout the study. The work relies on available data, excludes personal identifiers, aggregates results where appropriate, and documents methodological limitations. These decisions collectively support transparency, reproducibility, and critical scrutiny of the findings.

Versioning and Artefact Release. The observational measurement system is maintained in a version-controlled Git repository [Alk26]. All empirical results reported in this thesis were generated using release `v1.1-thesis`, whose tag identifies the exact code state used to produce the datasets, derived feature matrices, manifests, and validation reports described in the study.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.8 Contributions and Limitations

This section summarises the contribution scope and methodological limitations of the research design developed in this chapter. The detailed classification of the thesis outputs as DSR constructs, model, methods, instantiation, and methodology contribution is provided in Section 3.4. The present section therefore focuses on how the methodological design supports the research questions and where its limits remain.

3.8.1 Contribution Scope

Empirical contribution. The empirical contribution of the methodology is that it enables the quantification of longitudinal patterns of content availability and archival persistence for Syria-related YouTube videos across the observation period 2011–2025. For **RQ1**, the design supports measurement of externally observable platform availability states, archival presence, temporal coverage, right-censoring, and CDX temporal fragmentation. For **RQ2**, it supports statistical analysis of associations between taxonomy-derived content attributes and observable availability outcomes. The contribution is therefore not a causal explanation of content removal, but an empirically grounded account of what can and cannot be established from external platform and archival evidence.

Methodological contribution. The methodological contribution is a reproducible observational framework for studying platform-mediated digital memory under conditions of infrastructural instability and platform opacity. The framework combines platform-facing responses, Wayback archival traces, and curated corpus metadata while preserving the evidentiary limits of each source. It provides a repeatable procedure for sampling, live infrastructure interaction, semantic interpretation, archive-side evidence-state reconstruction, content-taxonomy-based feature construction, completeness operationalisation, and quantitative observational analysis. The methodological value of this framework lies in making partial traces, missingness, indeterminacy, and observability limits explicit rather than discarding or silently repairing them.

Practical and system-level contribution. The practical contribution is the derivation of system-level design implications from observed preservation gaps and infrastructural constraints. For **RQ3**, the methodology supports the evaluation of how observability flags, completeness indicators, feedback channels, and reproducibility instruments can make the limits of digital-memory preservation more visible. These implications are grounded in the observed behaviour of the measurement system and remain limited to what can be supported by external evidence.

3.8.2 Limitations

The research design remains observational. It cannot identify causal effects, platform intent, moderation rationale, or definitive reasons for content unavailability. All claims are restricted to externally observable platform-facing and archival signals.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The analysis is also conditioned by archival coverage. Videos without Wayback records, without replayable archived HTML, or without recoverable semantic signals cannot support the same level of reconstruction as videos with richer archival evidence. The completeness logic and explicit missingness handling mitigate this interpretive risk, but they do not remove the underlying observability limit.

The corpus and sample are defined by the available Syria-related YouTube identifiers and by the stratified sampling procedure described in Section 3.3.3. The results are therefore valid for the defined analytical population and should not be generalised to all YouTube content or to all conflict-related digital documentation without further study.

Field testing with the constituent community is not completed within the scope of this thesis. In DSR terms, this means that the Relevance Cycle does not fully close through deployment with organisations such as SyrianMemory or Mnemonic. Such deployment remains a future evaluation step.

3.8.3 Alignment of Methodological Elements with Research Questions

Table 3.3 summarises how the main methodological elements of the chapter support the three research questions.

Table 3.3: Alignment of methodological elements with research questions (RQ1–RQ3).

Methodological element	**Role in this thesis**	**RQs**
Structured literature grounding	Defines the constructs, variables, and methodological precedents used to design the observational framework.	All
Corpus definition and hash-based sampling	Defines the fixed analytical subset of Syria-related YouTube identifiers used for observation, feature construction, and analysis.	All
Semantic interpretation and evidence-state reconstruction	Turns Internet Archive evidence into archive-side evidence states, evidence-state sequences, event indicators, right-censoring markers, and CDX temporal-gap indicators.	RQ1
Human taxonomy feature construction	Turns curated taxonomy labels into fixed content-related variables used in the RQ2 regression analysis.	RQ2
Archival completeness construct	Measures preservation as a graded condition based on observable archival layers and supports interpretation of observability gaps.	RQ3
Quantitative observational analysis	Describes availability and archival patterns and tests associations between taxonomy-derived attributes and observable availability outcomes.	RQ1, RQ2

Continued on next page

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 3.3: Alignment of methodological elements with research questions (RQ1–RQ3) continued.

Methodological element	**Role in this thesis**	**RQs**
Design Science evaluation	Checks whether the measurement system satisfies the research-gap criteria: differentiated observability, partial evidence retention, reproducibility, and non-causal interpretation.	All

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 4

Empirical Analysis

4.1 Analytical Scope and Execution Context

This chapter reports the empirical observations produced by executing the system described in Chapter 3 as a methodological instrument. The analysis is strictly observational and documents externally observable patterns of availability, persistence, and absence without attributing causality, intent, or normative significance. All results presented in this chapter are derived from the data artefacts generated by the measurement procedure specified earlier. No additional filtering, transformation, or manual correction is introduced at the analysis stage.

The analytical findings reported in the sections that follow are therefore understood as descriptions of externally observable patterns associated with systems whose internal logic remains inaccessible.

4.2 Corpus Definition, Units, and Data State

4.2.1 Corpus Definition

The empirical corpus analysed in this chapter consists of YouTube video identifiers curated by a non-governmental organisation (NGO). The NGO-provided dataset serves exclusively as a corpus-defining input that determines the set of video identifiers included in the study. No analytical claims are derived from the NGO's selection criteria, ranking logic, metadata annotations, or selection and inclusion decisions. The NGO dataset is treated solely as an external source that specifies the study population.

4.2.2 Units of Analysis and Temporal Granularity

The primary unit of analysis is the individual video identifier. All observations are associated with specific video identifiers and, where applicable, with timestamped records obtained from web archive snapshots or live platform checks. Temporal granularity is defined by snapshot timestamps, which serve as the basis for reconstructing per-video temporal trajectories of availability and absence. Results are aggregated at the per-video level to enable comparison across repeated observations over time.

4.2.3 Data Sources and Completion Status

The empirical analysis reported in this chapter draws on multiple data sources collected through the observational measurement system described in Chapter 3. This subsection documents the data sources used and specifies their respective collection and completion status at the time the analysis was executed.

Three primary data sources are integrated.

First, Wayback CDX snapshot metadata were retrieved via the Internet Archive CDX API. These records provide snapshot-level information on archival capture attempts, including timestamps, HTTP status code, MIME types, and references to archived payloads. CDX snapshot metadata collection was completed and frozen prior to analysis and is treated as complete for the purposes of this chapter.

Second, YouTube platform availability checks were performed through live platform requests. These checks yield a single point-in-time, video-level observation of platform-facing availability status for each video identifier in the analytical corpus. All planned YouTube platform checks were executed successfully and are treated as complete and fixed at execution time.

Third, archived HTML snapshots were retrieved from the Internet Archive's Wayback Machine replay layer and constitute the frozen HTML evidence used throughout the analysis. While HTML snapshot coverage remains uneven across videos and timestamps, all analyses reported in this chapter and in subsequent chapters rely on this fixed and complete retrieval state.

As a consequence of these differing completion states, analyses reported in this chapter rely only on those data sources whose collection was complete at execution time. The completion status of each data source is made explicit to ensure transparency regarding the empirical basis of the analysis.

4.3 Wayback CDX Temporal Evidence

This section analyses temporal characteristics of the Wayback CDX snapshot metadata for the analysed corpus. All results are derived exclusively from CDX records and describe observable patterns of archival coverage without reference to HTML payloads or live platform responses.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.3.1 Snapshot Frequency and Archival Lifespan

Snapshot frequency varies substantially across the corpus. As summarised in Table 4.1, the median number of CDX snapshots per video is one, with higher quantiles indicating a strongly right-skewed distribution. Gap count and total gap duration statistics are computed only for videos with at least two CDX snapshots, as gaps are structurally undefined for single-snapshot videos; this reduces the effective sample from 18,170 to 5,449 for those measures. In proportional terms, exactly one snapshot is observed for 70.01% of videos, whereas 29.99% have at least two snapshots and 10.82% have at least five snapshots (Table 4.2).

Figure 4.1 visualises the highly uneven distribution of CDX snapshot counts per video, using a logarithmic frequency scale to accommodate the pronounced skew.

Table 4.1: Core descriptive statistics for Wayback CDX temporal measures.

Measure	n	min	p25	median	p75	mean
CDX snapshots per video	18,170	1.000	1.000	1.000	2.000	2.973
Archival lifespan (days)	18,170	0.000	0.000	0.000	2.121	447.307
CDX temporal gap count (> 1 day)	5,449	0.000	1.000	1.000	4.000	4.001
Total gap duration (days)	5,449	0.000	52.248	603.906	3,150.152	1,490.968

Archival lifespan, defined as the elapsed time between the earliest and latest CDX snapshot recorded for each video, mirrors this pattern. As shown in Table 4.1, the median archival lifespan is zero days, corresponding to videos represented by a single snapshot. At the same time, the upper tail extends over multiple years, with the 90th percentile exceeding 1,600 days and the maximum observed lifespan exceeding 5,200 days. Overall, 18.19% of videos exhibit an archival lifespan of at least one year (Table 4.2).

The distribution of non-zero archival lifespans exhibits a heavy-tailed structure, indicating that a small subset of videos remains archived over extended periods. This pattern is illustrated in Figure 4.2 and is reported as supplementary evidence in the appendix to avoid over-emphasising rare long-lived cases in the main analysis.

Table 4.2: Selected proportions describing Wayback CDX coverage and discontinuities.

Metric	Percentage (%)
Exactly one CDX snapshot per video	70.01
At least two CDX snapshots	29.99
At least five CDX snapshots	10.82
At least ten CDX snapshots	5.58
Zero-day archival lifespan	70.01
Archival lifespan of at least 365 days	18.19
Multi-snapshot videos with at least one CDX temporal gap (> 1 day)	84.35

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

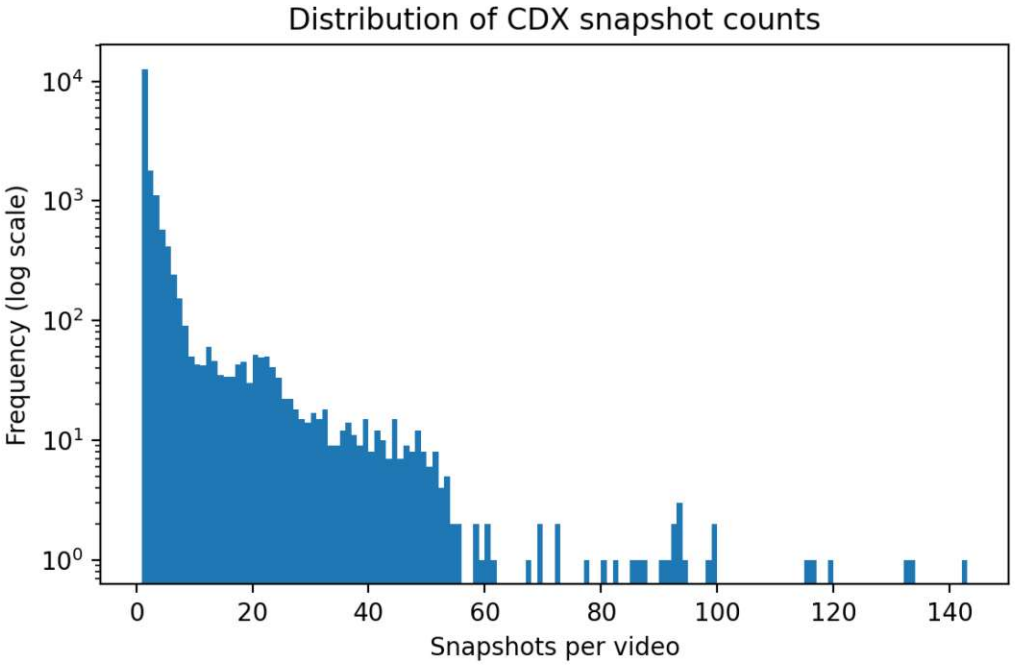

Figure 4.1: Distribution of CDX snapshot counts per video (log-scaled frequency).

4.3.2 CDX Temporal Gaps and Discontinuities

CDX temporal discontinuities were examined for videos with more than one CDX snapshot ($n = 5{,}449$). For each such video, intervals exceeding one day between consecutive CDX snapshots were identified and counted. As reported in Table 4.1, the median CDX temporal gap count is one, while higher quantiles indicate substantial variation across videos.

Figure 4.3 presents the distribution of CDX temporal gap counts using discrete bins. The majority of fragmented videos exhibit only a small number of gaps, while progressively fewer videos display more extensive discontinuities. This representation avoids over-emphasising rare extreme cases and highlights typical patterns of temporal fragmentation.

Cumulative gap duration further highlights the discontinuous nature of archival coverage. As shown in Table 4.1, the median cumulative gap duration exceeds 600 days, with upper quartiles extending into multiple years. These results indicate that repeated archival presence does not imply temporal continuity.

4.3.3 Terminal Snapshot Patterns

To characterise the cessation of archival traces, the year of the final observed CDX snapshot was extracted for each video. Terminal snapshots are distributed across a wide

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

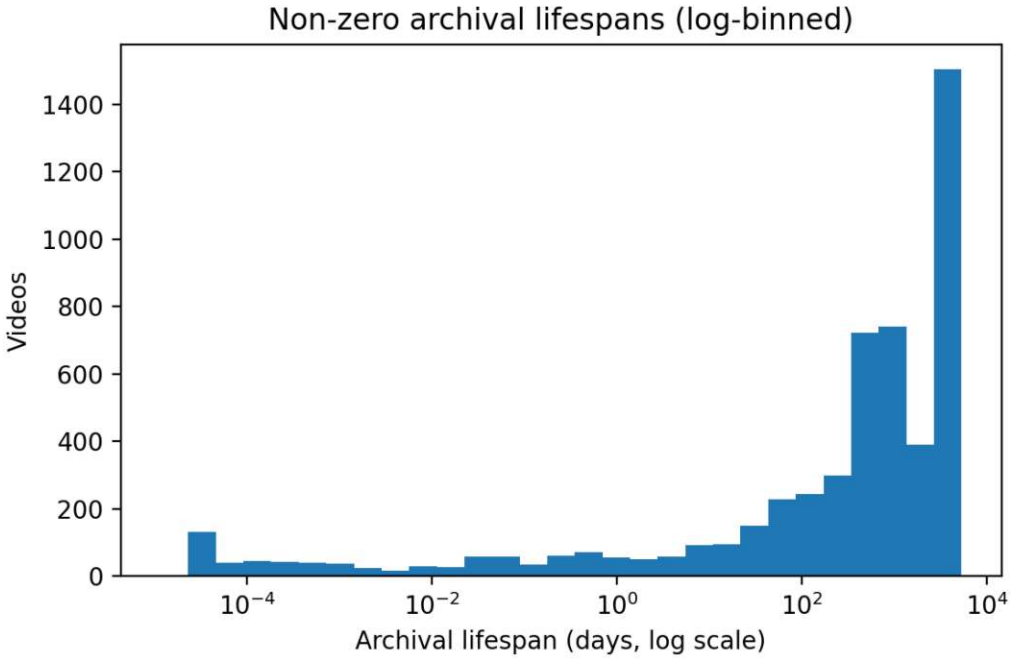

Figure 4.2: Distribution of non-zero archival lifespans in days (log-scaled x-axis).

temporal range, indicating that archival traces cease at different points in time across the corpus.

Figure 4.4 visualises the distribution of terminal snapshot years. The concentration of terminal observations in the most recent years reflects right-censoring due to the finite observational window rather than a sudden increase in content disappearance.

4.4 Wayback HTML Snapshot Evidence

This section examines the empirical properties of HTML snapshots retrieved from the Internet Archive's Wayback Machine for the analysed corpus. The analysis focuses on (i) HTML snapshot coverage, (ii) structural characteristics of archived pages, and (iii) the presence of YouTube-specific semantic signals. All results are strictly descriptive and document the empirical conditions under which archived HTML can support, or limit, platform observability. No claims regarding content removal, moderation intent, or temporal change are made in this section.

4.4.1 HTML Snapshot Coverage

Table 4.3 summarises HTML snapshot coverage relative to the CDX snapshot universe. A total of 28,602 HTML snapshots were retrieved, covering 18,117 distinct video identifiers.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

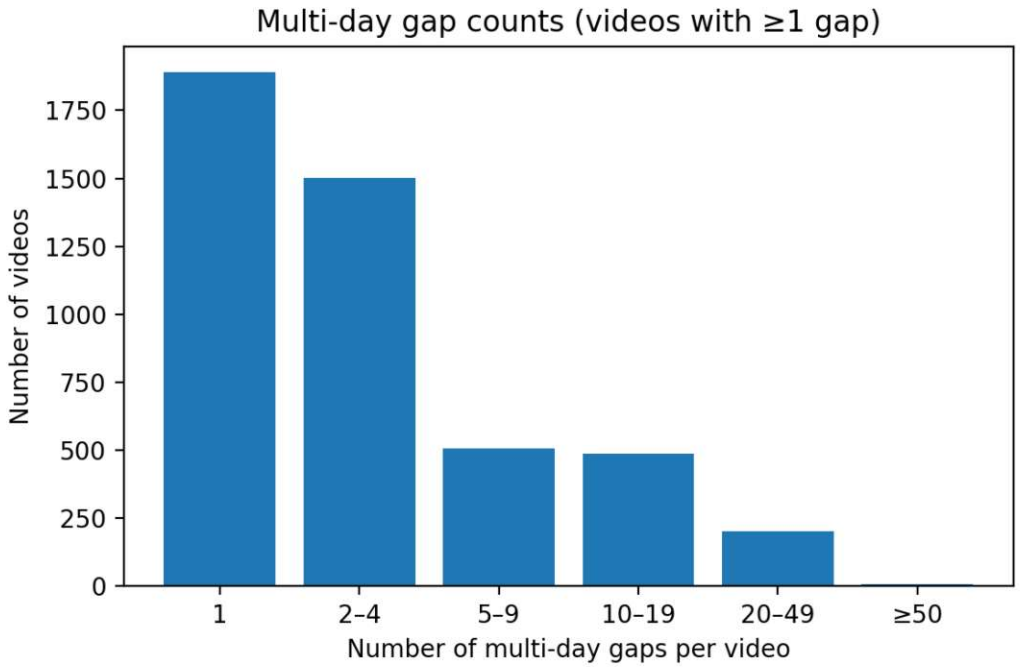

Figure 4.3: Distribution of multi-day gap counts per video for videos with at least one gap (> 1 day).

Table 4.3: Coverage of Wayback HTML snapshots relative to the CDX snapshot universe

Measure	**Count**	**Percentage**
HTML snapshots retrieved	28,602	–
Distinct videos with HTML	18,117	–
CDX snapshots (reference)	54,013	–
HTML snapshots relative to CDX	28,602	53.0%

Compared to 54,013 CDX snapshots, the retrieved HTML snapshots constitute an incomplete subset of the CDX snapshot universe.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

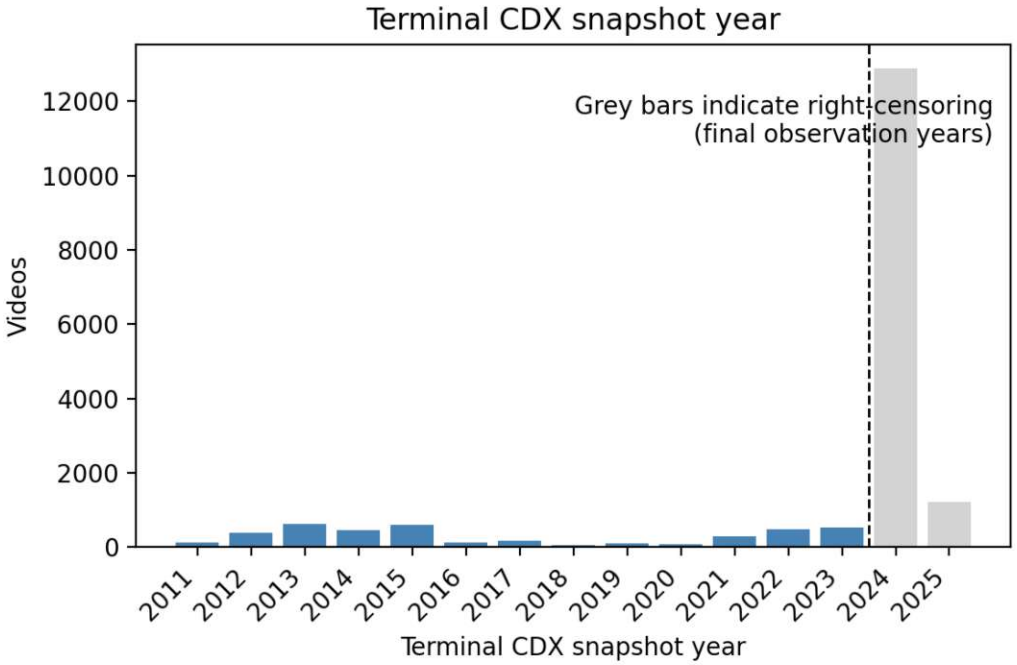

Figure 4.4: Year of last observed CDX snapshot per video. Grey bars indicate right-censored final observation years.

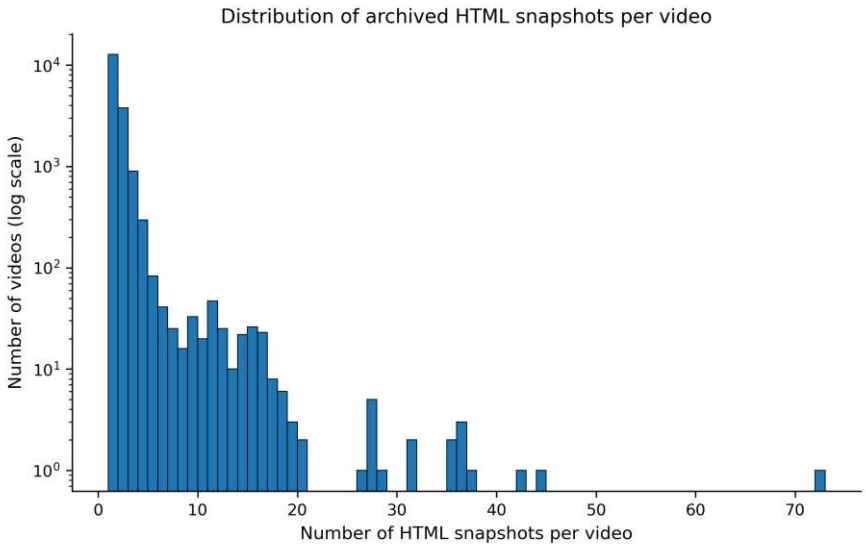

Figure 4.5: Distribution of archived HTML snapshots per video. The distribution is strongly right-skewed: most videos are represented by a single snapshot, while a small subset exhibits substantially higher snapshot counts.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Figure 4.5 indicates strongly uneven HTML coverage across videos, with a pronounced concentration at low snapshot counts and a long right tail. This heterogeneity indicates that HTML-level observability is unevenly distributed across the corpus and cannot be assumed to be uniform at the video level.

4.4.2 Structural Characteristics of Archived HTML

Table 4.4: Structural characteristics of archived HTML snapshots

Structural element	Snapshots	Percentage
Total HTML snapshots	28,602	100.0%
`<!DOCTYPE>` present	28,594	99.97%
`<html>` tag present	28,587	99.95%
`<head>` tag present	28,599	99.99%
`<body>` tag present	28,586	99.94%
Fully structured HTML	28,567	99.88%

As shown in Table 4.4, the majority of archived HTML snapshots exhibit complete canonical document structure. More than 99% of snapshots contain all core HTML elements, indicating that Wayback replay generally yields structurally valid HTML suitable for systematic parsing and inspection. A small residual fraction of snapshots lacks one or more canonical elements, consistent with isolated replay-layer or capture artefacts rather than systematic parsing failure.

Table 4.5: Composition of archived HTML snapshots by player response

Response shape	Snapshots	Percentage
Captured YouTube HTML	17,757	62.1%
Captured YouTube HTML (player missing)	5,605	19.6%
Captured non-YouTube HTML	3,517	12.3%
Wayback rate-limit notice	1,062	3.7%
Captured site error page	484	1.7%
Wayback robots exclusion	175	0.6%
Indeterminate HTML	2	<0.1%

Table 4.5 further shows that structurally valid HTML is returned across a range of response shapes. In addition to fully captured YouTube pages, the corpus includes Wayback-generated interstitials and non-YouTube responses. Structurally valid HTML is frequently present even for non-success HTTP responses, reflecting the wrapper-based nature of Wayback replay.

4.4.3 Playability Signals and Metadata Presence

As summarised in Table 4.6, YouTube playability signals are present in only 62.1% of archived HTML snapshots. Where present, the dominant playability status is `OK`, while a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 4.6: Presence of YouTube playability signals in archived HTML snapshots

Playability signal	**Snapshots**	**Percentage**
Total HTML snapshots	28,602	100.0%
With player response detected	17,757	62.1%
Without player response	10,845	37.9%
Playability status = `OK`	15,548	54.4%
Playability status = `LOGIN_REQUIRED`	1,285	4.5%
Playability status = `NULL`	0	0.0%
Playability status = `ERROR`	210	0.7%
Other playability states	714	2.5%

smaller subset indicates access restrictions or unresolved states. These signals are treated strictly as indicators of semantic presence within archived HTML and are not interpreted as evidence of availability or unavailability.

Despite high structural completeness, YouTube-specific semantic yield is partial: only 17,757 of 28,602 HTML snapshots (62.1%) expose a detectable player response (Tables 4.5 and 4.6).

Table 4.7: Presence of video-related metadata fields in archived HTML snapshots

Metadata field	**Snapshots**	**Percentage**
HTML title tag present	28,586	99.94%
Channel ID present	17,183	60.1%
Owner channel name present	16,241	56.8%
Short description present	16,276	56.9%
Video details title present	16,276	56.9%
All metadata fields present	16,241	56.8%

Table 4.7 shows substantial metadata attrition at the video-context level: 39.9% of snapshots lack a channel ID and 43.2% lack an owner channel name, while only 56.8% contain all listed metadata fields. As a result, the existence of an archived HTML snapshot does not guarantee semantic observability of video-level context.

Taken together, these findings show that archived HTML frequently preserves structural form while omitting or obscuring video-specific meaning. HTML snapshot evidence therefore defines an empirical boundary for platform observability, complementing but not substituting endpoint- or CDX-based evidence analysed in other sections.

4.5 YouTube Platform Availability States

This section reports empirical observations of video availability as observed through YouTube platform checks. The analysis is based exclusively on externally observable

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

platform responses collected during the execution of the measurement system described in Chapter 3. All results presented here are strictly descriptive and document point-in-time platform-facing outcomes without attributing intent, causality, or moderation rationale.

Availability states in this section are derived solely from endpoint-level responses obtained in *YouTube mode*. HTML payloads retrieved during *YouTube mode* requests are treated only as transport-level artefacts and are not used to infer availability semantics. Archival evidence and longitudinal observability are analysed separately in later sections.

4.5.1 Observed Availability Categories

For each video identifier, the system classified the platform response into a discrete availability category using externally observable signals from the endpoint. Each observation reflects the platform-facing state associated with a specific video identifier at the time of measurement.

Table 4.8 summarises the absolute and relative frequency of the observed availability categories across the analysed corpus.

Table 4.8: Distribution of YouTube platform availability states observed at the time of measurement

Availability state	**Video count**	**Proportion (%)**
Available	22,987	45.95
Error	22,451	44.88
Age-restricted	2,175	4.35
Unknown	1,321	2.64
Account terminated	1,088	2.18
Total	50,022	100.00

The largest share of observations corresponds to videos classified as *available* (45.95%), followed closely by responses classified as *error* (44.88%). Together, these two categories account for the majority of platform observations. Smaller proportions are associated with *age-restricted* content (4.35%) and videos linked to *terminated accounts* (2.18%).

A residual category labelled *unknown* accounts for 2.64% of observations. This category does not represent a platform-declared availability state but captures cases in which platform-facing availability could not be determined at the time of measurement. The nature of this category is specified below.

These distributional findings describe the platform-facing accessibility of a corpus of Syria-related audiovisual documentation whose relevance for collective memory formation is established in Chapter 2. Three of the four conditions of digital remembrance identified in that chapter — technical accessibility, institutional stability, and socio-cultural circulation — are reflected, at the level of externally observable platform signals, in the distributional patterns reported in Table 4.8; the temporal dimension of remembrance is addressed

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

through the longitudinal analysis in Section 4.6. The analysis does not attribute any particular meaning to the distribution of availability states: it documents them as externally observable conditions whose relationship to the theoretical constructs developed in Chapter 2 is a matter for interpretive analysis in Chapter 6.

4.5.2 Measurement Boundary: Rate-Limited Observations

All observations classified as *unknown* exhibit identical transport-level characteristics. Specifically, all such cases correspond to unsuccessful requests, return HTTP status code 429, and are associated with the error type `rate_limited`. These characteristics are consistent with request blocking due to detected request volume or traffic patterns rather than the provision of a semantic availability response.

Table 4.9: Transport-level characteristics of platform observations classified as `unknown`

Attribute	Observed value	Count
Request success	False	1,321
HTTP status code	429	1,321
Error type	`rate_limited`	1,321

Although HTML responses were returned in these cases, the responses do not contain standard video page content. Instead, they consist of platform-generated interstitial pages associated with rate-limiting enforcement and do not expose video availability semantics. Consequently, observations in the *unknown* category are treated as *unobservable platform states* rather than as evidence of availability, restriction, or removal.

These observations therefore delineate a measurement boundary at the platform layer. They document limits of platform observability at the time of data collection and are not interpreted as content-level outcomes.

4.6 Observed Availability Histories from Archived Evidence

This section examines longitudinal availability patterns of the analysed videos as observed through repeated YouTube HTML snapshots. The analysis is based on availability descriptors derived from CDX interaction records and selectively fetched HTML snapshots, following the methodology defined in Chapter 3. The objective is not to infer platform intent or moderation rationale, but to characterise the empirical conditions under which availability and disappearance can be observed over time. The descriptors analysed here are treated as derived indicators of longitudinal observability based on repeated external measurements, rather than as behavioural or causal representations of platform-side processes.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.6.1 History coverage and corpus-level completeness

Table 4.10 summarises the coverage of availability descriptors across the analysed corpus. Availability descriptors were constructed for all videos with at least one CDX snapshot (n = 18,170). Each video appears exactly once in the trajectory table, ensuring full coverage of the CDX-defined corpus. The resulting dataset constitutes a structurally complete observational layer over the archival universe analysed in Section 4.3.

4.6.2 Semantic observability of availability states

Table 4.10 summarises the distribution of semantically observable states across the corpus. At least one semantically classifiable availability state (available, restricted, or unavailable) was observed for 14,835 videos. The remaining videos contain only CDX-level interaction evidence without recoverable semantic signals, that is, HTML snapshots lacking parsable player or error-state information. Among videos with semantic observability, availability is the dominant semantically observed state: 14,037 videos exhibit at least one observed `AVAILABLE` state. Observed restriction states are rare (five cases), and no descriptors exhibit an explicitly classified erosion state under the applied classification rules.

Observed unavailability is present in 1,215 descriptors, corresponding to 6.7% of the corpus. These cases represent videos for which a terminal unavailable state was directly observed in the HTML-derived evidence. The absence of an observed unavailable state does not imply continued availability, but reflects the limits of observation within the available snapshot data.

Table 4.10: Video-level summary of archived availability descriptors derived from Wayback evidence.

Metric	Videos
Total videos with descriptors	18,170
Videos with any observed semantic state	14,835
Videos with observed availability	14,037
Videos with observed unavailability	1,215
Right-censored descriptor	16,955
Videos with at least one CDX temporal gap ($>$ 1 day)	4,596

4.6.3 Snapshot-Level Classifier-Label Distribution

In addition to trajectory-level descriptors, the classification of individual archived observations provides a finer-grained view of the implemented snapshot-level classifier output. Figure 4.6 shows the distribution of classifier labels across all snapshot observations. Five labels are present: `unknown`, `available`, `redirect`, `unavailable`, and `restricted`. The label `unknown` records archive-side evidence that cannot be assigned to one of the seven interpretable archive-side evidence states. The distribution is highly

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

uneven. The majority of snapshots are classified as `unknown`, followed by `available` observations. Redirect-based interactions form a smaller but non-negligible share of observations, while explicitly observed `unavailable` labels occur less frequently. Restricted labels are only marginally represented. This distribution describes the empirical composition of classified observations at the snapshot level.

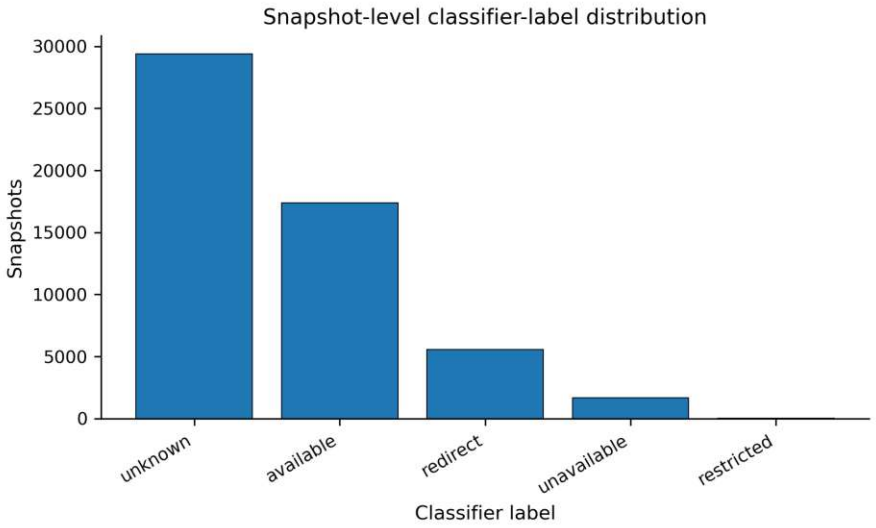

Figure 4.6: Distribution of implemented snapshot-level classifier labels across all snapshot observations.

4.6.4 Observed transition counts per video

Figure 4.7 shows the distribution of transition counts per video. A transition is defined as a change in snapshot-level classifier label between two adjacent observations. The distribution is concentrated at zero, indicating that most videos do not exhibit observable label changes across recorded snapshots. A smaller subset of videos exhibits one or more transitions.

4.6.5 CDX Temporal Discontinuities

Table 4.10 summarises the occurrence of CDX temporal gaps across descriptors. CDX temporal gaps exceeding the implemented one-day threshold were identified in 4,596 descriptors (25.3% of videos). These gaps indicate intervals between consecutive CDX observations during which no additional CDX snapshot is recorded. Their presence shows that repeated archival observations do not provide continuous archive-side coverage over time. Gap durations correspond to intervals between CDX timestamps. They indicate missing observations between captures and do not represent how long a video remained in a specific archive-side evidence state.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

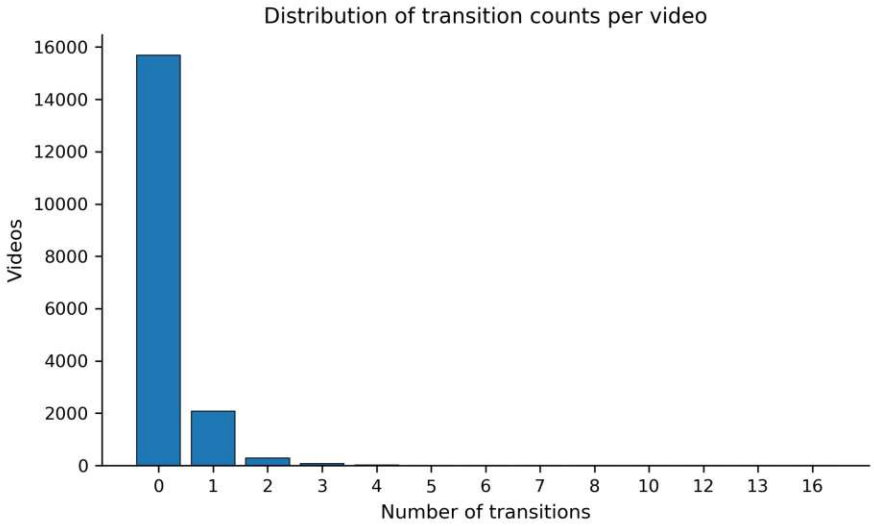

Figure 4.7: Distribution of observed snapshot-level classifier-label transition counts per video.

4.6.6 Gap Structure Across Snapshot-Level Classifier Labels

Figure 4.8 shows the proportion of snapshot observations preceded by a CDX temporal gap, that is, cases in which the interval to the previous observation exceeds the implemented one-day gap threshold, by snapshot-level classifier label. Restricted labels exhibit the highest proportion of preceding gaps, followed by redirect and unknown labels. Available labels show the lowest proportion of gap-preceded observations, while unavailable labels occupy an intermediate position. This distribution describes how CDX temporal discontinuities are distributed across the implemented snapshot-level classifier output.

4.6.7 Right-censoring and terminal uncertainty

Figure 4.9 shows the distribution of terminal descriptor outcomes across the corpus. The majority of descriptors (16,955 videos, 93.3%) are right-censored, meaning that their final observed state does not indicate unavailability. In these cases, disappearance cannot be empirically confirmed within the observation window. The figure shows the asymmetry between observed unavailability and right-censored descriptors.

For the majority of videos, the empirical record does not contain an observed terminal unavailable state within the available observation window. Accordingly, the record does not distinguish within this window between continued availability and disappearance without observable trace. This describes a property of the observational record produced by the applied measurement procedure. The implications of this result are addressed in Chapter 6.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

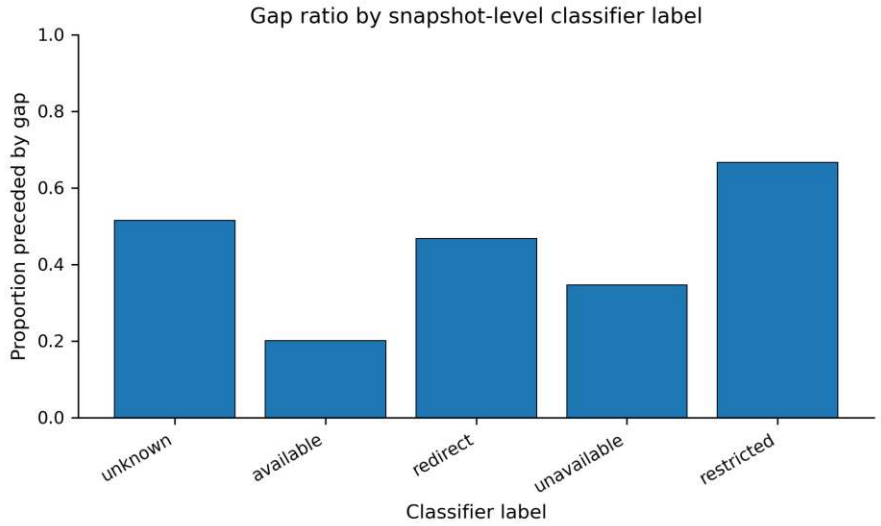

Figure 4.8: Proportion of snapshot observations preceded by a CDX temporal gap, by snapshot-level classifier label.

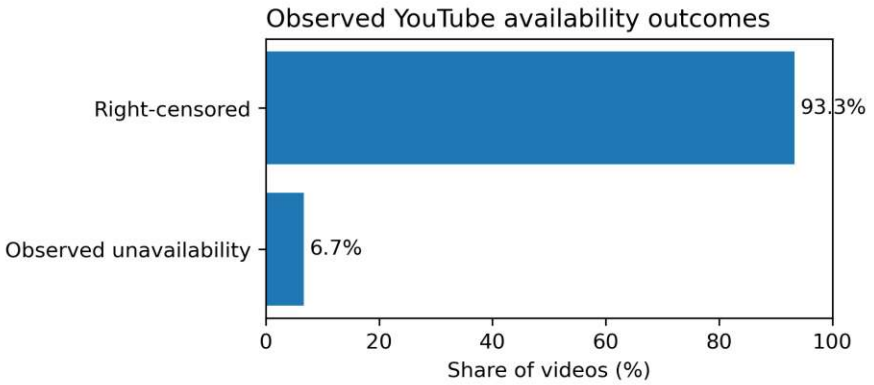

Figure 4.9: Observed YouTube availability descriptor outcomes. The figure contrasts videos with an explicitly observed terminal unavailable state and descriptors that remain right-censored at the end of the observation period.

4.6.8 Interpretive scope

Taken together, these results describe YouTube availability observability in terms of semantic observability, right-censoring, and temporal discontinuities. The contribution of this subsection is to characterise the empirical properties of the observed record over time, rather than to provide causal explanations for content removal. These results complement the archival perspective provided by the Wayback CDX analysis and provide the empirical basis for the discussion in Chapter 6.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.7 Cross-Source Observability Boundaries

This section examines the empirical boundaries of joint observability across the three data sources used in this study: Wayback CDX records, archived HTML snapshots, and live YouTube platform responses. The analysis documents where signals co-occur, where they diverge, and where observability breaks down due to infrastructural or replay-layer constraints. All results are strictly descriptive and concern signal presence or absence across sources. No claims regarding temporal change, persistence, or disappearance are made in this section.

Unless stated otherwise, counts are reported at the *snapshot level* for archival data and at the *video level* for live platform checks.[1]

4.7.1 CDX and HTML Observability Overlap

The Wayback CDX index defines the maximal archival universe for the analysed corpus, while archived HTML snapshots represent a selectively retrieved subset of this universe. Table 4.11 quantifies the overlap between CDX records and replayable HTML snapshots at identical timestamps.

Table 4.11: Intersection between Wayback CDX records and archived HTML snapshots.

Measure	Count
Total CDX snapshot records	54,013
HTML snapshots retrieved	28,602
Exact CDX–HTML timestamp matches	28,602
CDX snapshots without HTML replay	25,411

Although HTML snapshots were retrieved for nearly all videos with CDX presence (18,117 of 18,170 videos), Table 4.11 shows that approximately 53% of CDX snapshots have a corresponding replayable HTML capture at the same timestamp. This demonstrates that CDX-level observability and HTML-level observability constitute distinct empirical layers. CDX presence alone does not imply the availability of page-level semantic evidence at the same archival timestamps.

HTML snapshot evidence is near-complete at the video level: 18,117 of 18,170 videos with CDX presence have at least one replayable HTML snapshot, but incomplete at the timestamp level: only 28,602 of 54,013 CDX snapshot records (53.0%) have a corresponding replayable HTML capture at the same timestamp (Tables 4.3 and 4.11).

4.7.2 Semantic yield (archived html) Conditional on CDX Presence

Archived HTML snapshots vary substantially in their capacity to expose YouTube-specific semantic signals. Table 4.12 reports the semantic yield of archived HTML snapshots

[1] Snapshot-level observations refer to individual CDX or HTML capture timestamps. Video-level observations refer to a single live YouTube platform check per video identifier.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

conditional on exact CDX–HTML timestamp matches.

Table 4.12: Semantic yield of archived HTML snapshots with exact CDX timestamp matches.

Response shape	Player response present	Snapshots
Captured YouTube HTML	Yes	17,757
Captured YouTube HTML (player missing)	No	5,605
Captured non-YouTube HTML	No	3,517
Wayback rate-limit notice	No	1,062
Captured site error page	No	484
Wayback robots exclusion	No	175
Indeterminate HTML	No	2

Only 62.1% of replayable HTML snapshots expose a YouTube player response. The remaining snapshots consist of structurally valid HTML that lacks video-level semantic signals due to replay-layer interstitials, missing player responses, or non-YouTube captures. Accordingly, the existence of a replayable HTML snapshot does not guarantee semantic observability of platform availability.

4.7.3 HTML and YouTube Signal Co-Presence

To assess cross-source alignment between archived HTML semantics and live YouTube platform states, the latest archived HTML snapshot at or before each live platform check was identified per video. Table 4.13 reports the joint distribution of live YouTube endpoint-derived availability states and HTML-derived semantic states.[2]

While Table 4.13 provides exact counts, the distribution is highly skewed across orders of magnitude. To reveal structural asymmetries that are not immediately apparent from the aggregated table, Figure 4.10 visualises a more granular rendering of the same underlying HTML–YouTube pairing using logarithmic scaling.

The figure shows that semantic alignment is uneven across platform states. In particular, large numbers of videos classified as *available* or *error* at the live platform layer lack any corresponding HTML semantic signal, indicating the structural limits of cross-source observability.

[2]Unit of analysis: video identifier. Each video contributes at most one HTML–YouTube pair, based on the latest HTML snapshot preceding the live check.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 4.13: Joint distribution of live YouTube platform states and archived HTML semantic states.

YouTube live state	HTML semantic state	Videos
Available	OK	6,769
Available	No player response	1,450
Available	Login required	395
Available	Other / none	147
Error	OK	6,383
Error	No player response	1,378
Error	Login required	444
Error	Other / none	121
Age-restricted	No player response	219
Age-restricted	Login required / OK / none	188
Account terminated	No player response	102
Account terminated	Any semantic state	175
Unknown	Any semantic state	448

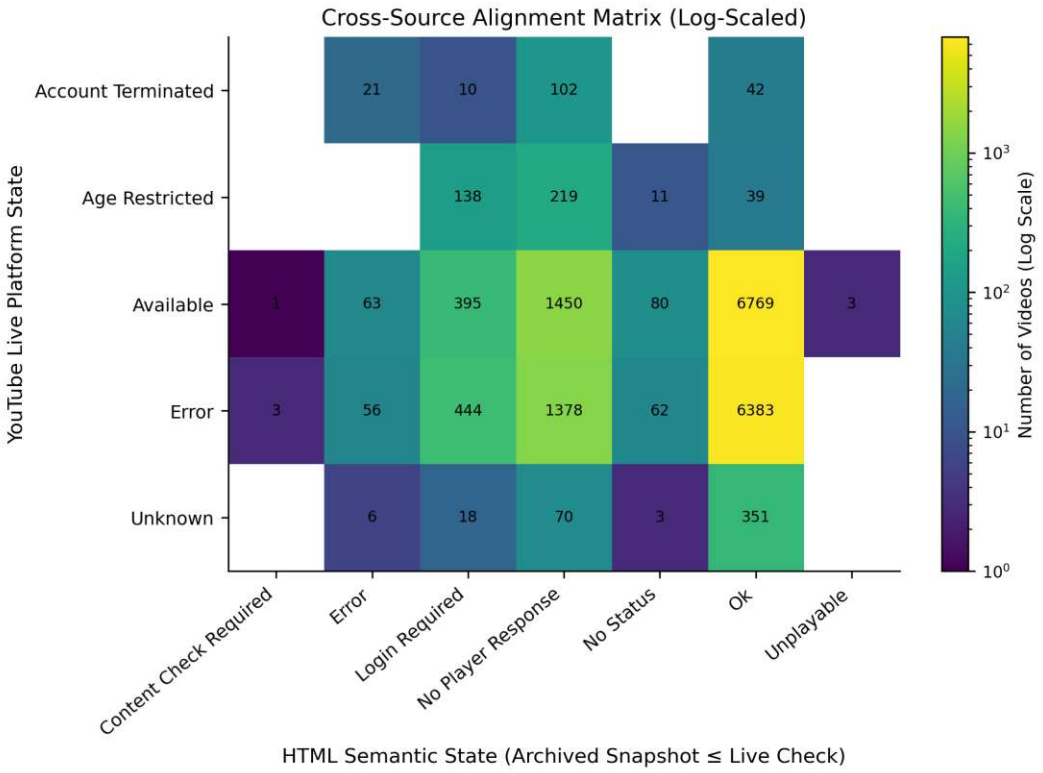

Figure 4.10: Log-scaled cross-source alignment matrix between live YouTube platform states and archived HTML semantic states. Cells indicate the number of videos for each state combination. Logarithmic scaling reveals sparse but systematic misalignments that are visually suppressed in linear representations.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.7.4 Structural Observability Boundaries

Finally, Table 4.14 quantifies the proportion of videos for which at least one archived HTML snapshot with recoverable YouTube semantic evidence was observed prior to the live YouTube platform check.

Table 4.14: Videos with at least one prior archival HTML semantic signal before live YouTube checks.

YouTube live state	Videos	With HTML semantics	Proportion
Available	22,987	7,379	0.321
Error	22,451	6,992	0.311
Age-restricted	2,175	189	0.087
Account terminated	1,088	75	0.069
Unknown	1,321	383	0.290

Figure 4.11 visualises, relative to Table 4.14, the complementary proportion of videos for which no archival HTML semantic signal was observable in any Wayback HTML snapshot prior to the live YouTube platform check, grouped by live platform state.

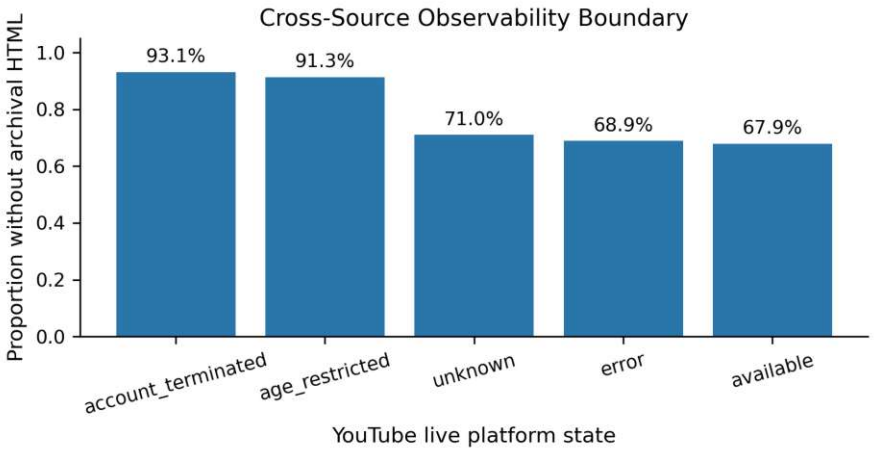

Figure 4.11: Videos without prior archival HTML semantic evidence before live YouTube checks.

Taken together, these results document uneven archival HTML semantic observability across live platform-facing availability states. They identify where cross-source comparison is supported by prior semantic archival evidence and where it is structurally limited by its absence. These observability boundaries condition, but do not themselves explain, platform-mediated availability outcomes.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.8 Content Taxonomy Coverage and Feature Availability

This section reports descriptive properties of the human taxonomy features constructed for regression analyses addressing **RQ2**. All results are derived from the frozen per-video feature dataset generated during the preprocessing stage described in Chapter 3. The purpose of this section is to document (i) the coverage of human-assigned taxonomy labels, (ii) the realised reduction of those labels into grouped analytical categories, (iii) the transformation of grouped categories into regression-ready semantic-top identifiers, and (iv) the axis-level availability of categorical covariates. No interpretive, semantic, or causal claims are made.

4.8.1 Human Taxonomy Label Coverage

Table 4.15 reports the distribution of videos with and without human-assigned taxonomy labels across the analytical corpus.

Table 4.15: Human taxonomy label coverage across the analytical corpus

Has human label	Videos	Proportion
False	371	0.007
True	49,651	0.993

The large majority of videos (99.3%) contain at least one human-assigned taxonomy label on one or more of the **RQ2** axes. Videos without labels are retained explicitly and treated as structurally unlabelled cases rather than missing observations.

4.8.2 Taxonomy Reduction Across Axes

This subsection documents the realised reduction from field-specific raw taxonomy values to retained grouped categories in the frozen preprocessing output. The purpose is purely descriptive. Table 4.16 reports how many distinct raw values were observed in each of the three taxonomy fields used for **RQ2**, how many semantic groups were created through manual many-to-one mapping, and how many grouped categories were retained after admissibility filtering. These counts describe the realised feature-construction output on the analytical corpus and do not introduce additional modelling assumptions beyond those already defined in Chapter 3.

Table 4.16: Reduction of field-specific taxonomy values into retained grouped categories

Axis	Source field	Raw values	Semantic groups	Retained groups
Axis A	`main_classification`	29	9	7
Axis B	`group`	11	6	5
Axis C	`subcategory`	114	13	5
Total	—	**154**	**28**	**17**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The table shows that the construct operates on three axis-specific source fields rather than on a single pooled taxonomy. The observed raw value spaces are reduced independently within each axis through manual semantic grouping and then further restricted through admissibility filtering before final per-video feature assignment.

4.8.3 Structural Consistency Between Labels and Semantic-Top Features

Semantic-top features are derived from human-assigned taxonomy labels through a two-step transformation process. First, each raw field value is mapped to a manually defined *semantic core*, representing a canonical higher-level content category within its axis. Second, admissibility filtering determines which grouped categories remain available for final semantic-top assignment as categorical predictors in the regression models.

This transformation does not rely on natural language processing, clustering, or statistical inference. All semantic cores are defined *ex ante* and applied via fixed lookup tables. NLP outputs are used exclusively during preprocessing as diagnostic instruments to assess label dispersion and stability, and do not contribute directly to feature construction.

Table 4.17 reports the number of videos for which human labels are present but no admissible semantic-top feature could be assigned on any axis. These cases arise solely from fixed admissibility constraints defined during feature construction and do not reflect missing data or post hoc filtering.

Table 4.17: Structural consistency between human labels and semantic-top features

Condition	**Videos**
Labels present but no semantic top	16
Videos retained after constraint	50,022

All reported counts follow directly from the preprocessing rules defined in Chapter 3. No rows are removed or altered at the analysis stage.

4.8.4 Axis-Level Feature Availability

This subsection reports the distribution of semantic-top identifiers per axis. Semantic tops are treated strictly as categorical identifiers derived from human labels. They do not represent topics, themes, or inferred meanings, and they are not the output of NLP models.

Table 4.18 lists the retained grouped categories that remain available for semantic-top assignment on each axis after admissibility filtering. These categories define the axis-specific categorical feature space documented in the tables below.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 4.18: Retained grouped categories by taxonomy axis

Axis	Source field	Retained grouped categories
A	`main_classification`	Armed violence and combat; Protest and civic mobilisation; Human casualties and losses; Repression and violations; Displacement and humanitarian crises; Humanitarian activity; Political activity
B	`group`	Civilian; Military; Martyrs and fatalities; Humanitarian; Violations
C	`subcategory`	Human casualties and losses; Armed violence and combat; Organisational or military activity; Protest and civic mobilisation; Repression and violations

4.8.5 Cross-Axis Feature Presence

Table 4.19 summarises how many axes contain at least one admissible semantic-top assignment per video. Counts reflect feature availability after semantic-core mapping and admissibility checks.

Table 4.19: Cross-axis semantic-top presence per video

Feature presence	Videos	Proportion
Labels on two axes	29,380	0.587
Labels on all three axes	12,719	0.254
Labels on one axis	7,536	0.151
No semantic-top features	387	0.008

This distribution documents the uneven availability of content-related covariates across axes and makes explicit the small subset of videos without derived semantic-top features.

Position in chapter structure. This section completes the descriptive documentation of content-related covariates and directly precedes the feature freeze described in Section 4.9. All reported variables constitute fixed analytical inputs to the statistical models presented in Chapter 5.

4.9 Feature Freeze and Statistical Input Dataset

From the frozen availability trajectory, a structured feature dataset was generated in accordance with the feature definitions specified in Chapter 3. The dataset consists of one row per video identifier and includes time-to-event variables, evidence-state duration measures, archival-quality indicators, and content-related covariates. All variables are computed without ad hoc transformation, and missing values are explicitly encoded.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The resulting feature dataset is frozen at this point and is not analysed further in the present chapter. It constitutes the sole analytical input to the statistical models reported in Chapter 5.

4.10 Summary and Analytical Transition

This chapter has documented the empirically observable properties of availability-related signals across three distinct measurement layers: Wayback CDX records, archived HTML snapshots, and live YouTube platform responses. Across these layers, the analysis has established systematic patterns of sparsity, discontinuity, and partial overlap. CDX records provide broad but temporally fragmented archival traces; HTML snapshots preserve structural form but only selectively expose platform semantics; and live YouTube checks yield point-in-time availability states that are themselves subject to measurement boundaries. Taken together, these results indicate that none of the individual data sources provides a complete representation of video availability.

The chapter has also delineated the empirical boundaries of joint observability across sources. Cross-source non-alignment is consistent with infrastructural, replay-layer, and platform-level constraints. As a result, absence of evidence at one layer cannot be interpreted as evidence of disappearance without further contextualisation. These findings establish the empirical basis for the interpretive discussion in Chapter 6.

The chapter also documents the derived archival availability histories and the availability of taxonomy-based content features. Together, these outputs form the frozen input to the statistical analyses reported in Chapter 5.

Chapter 5 builds on this empirical foundation by applying statistical analyses to the frozen feature datasets, deriving the system-level design response for **RQ3**, and evaluating the observational measurement system against the criteria defined in Chapter 3. Chapter 6 then synthesises the empirical and statistical findings in relation to digital memory, platform-mediated forgetting, and the limits of archival evidence.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

Statistical Analysis and System Evaluation

5.1 Purpose and Analytical Scope

This chapter applies statistical and design-evaluation procedures to analytically prepared datasets derived from the observational measurement system specified in Chapter 3 and empirically characterised in Chapter 4. The chapter has four functions. First, it addresses **RQ1** by analysing observable patterns of availability loss and archival observability using time-to-event feature representations applied to an availability-trajectory feature matrix. Second, it addresses **RQ2** by testing statistical associations between final observed availability outcomes and content-level attributes using a separate content feature matrix. Third, it addresses **RQ3** by formulating a system-level design response to the observability limits documented in Chapter 4 and in the preceding RQ1 and RQ2 analyses of this chapter. Fourth, it evaluates the observational measurement system as the methodological artefact that supports all three research questions, using the criteria defined in Section 3.6.

All analyses are strictly observational. No causal interpretation, attribution of platform intent, or inference about moderation policy is implied. Model outputs are interpreted as conditional statistical associations constrained by the limits of external observability established in Chapter 4.

5.2 Analytical Datasets and Feature Matrices

This section documents the analytical datasets and system outputs used in this chapter. Two feature matrices are used for the statistical analyses of RQ1 and RQ2, while

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

observability artefact outputs are used for the RQ3 design response and the Design Cycle evaluation.

5.2.1 RQ1 Feature Matrix: Availability Trajectory Dataset

The feature matrix used to address **RQ1** is an availability trajectory feature matrix. Each row corresponds to one YouTube video identifier. The matrix records the platform-facing availability state for the full RQ1 population and adds archival trajectory variables where Wayback evidence is available. The purpose of this matrix is to support descriptive analysis of platform availability, archival observability, observed archival unavailability, and right-censoring.

The unit of analysis is the individual video identifier. Variables included in the RQ1 feature matrix fall into four classes: (i) temporal outcome variables capturing time to first observed unavailability and right-censoring, (ii) archival availability indicators derived from rule-based HTML-level parsing, (iii) archival trajectory variables reconstructed from Wayback Machine evidence, and (iv) archival coverage indicators denoting the presence or absence of any archival history.

All 50,022 video identifiers in the NGO-defined corpus are included in the RQ1 feature matrix. Observed archival unavailability constitutes the event of interest. For videos with archival history, absence of observed archival unavailability within the observation window results in right-censoring at the final observed archival timestamp. Videos without archival history are retained and represented as structurally unobserved with respect to archival availability transitions.

Table 5.1 summarises the corpus composition, observed platform unavailability, and archival observability underlying the RQ1 descriptive analysis of platform availability and archival observability.

Table 5.1: RQ1 dataset summary: corpus size, platform availability, and archival observability.

Metric	Videos
Total videos analysed (RQ1 universe)	50,022
Videos currently unavailable on YouTube	20,770
Videos with any Wayback history	18,170
Videos with observed Wayback unavailability	1,215
Right-censoring videos (Wayback history, no observed unavailability)	16,955

5.2.2 RQ2 Feature Matrix: Content Attribute Dataset

The feature matrix used to address **RQ2** is a content attribute feature matrix. Each row corresponds to one YouTube video identifier with at least one admissible semantic-top assignment. The purpose of this matrix is to support cross-sectional regression analysis of final platform-observed semantic availability outcomes.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The unit of analysis is the individual video identifier. The dependent variable is derived at analysis time from the latest observable YouTube platform response and encodes final semantic availability (`playable` versus `explicitly_unavailable`). This outcome variable is not stored in the feature matrix itself.

Independent variables consist exclusively of content-level attributes derived from human-assigned taxonomy labels and their aggregation into semantic-top categories, as defined in Chapter 3. Each video included in the RQ2 content attribute feature matrix must contain at least one admissible semantic-top assignment on any taxonomy axis. Videos without an admissible semantic-top assignment on all three axes are excluded from the RQ2 feature matrix. No temporal or archival trajectory variables are included in the RQ2 feature matrix.

Table 5.2 reports the inclusion and exclusion counts resulting from the semantic-top feature availability constraint applied to the RQ2 feature matrix.

Table 5.2: RQ2 content-feature availability before outcome filtering

Stage	**Videos**
Videos in the RQ2 feature-construction universe	50,022
Excluded: no admissible semantic-top on any axis	387
Videos retained in the RQ2 content attribute feature matrix	49,635

Note. A video is excluded from the RQ2 content attribute feature matrix when all three semantic-top fields (`axis_A_main`, `axis_B_group`, `axis_C_subcat`) are missing after semantic-core mapping and admissibility filtering. This group includes structurally unlabelled videos and labelled videos that do not yield an admissible semantic-top assignment.

5.2.3 RQ3 and Evaluation Outputs: Observability Artefact Dataset

The outputs used for **RQ3** and for the subsequent Design Cycle evaluation are derived from the same frozen feature dataset described in Section 4.9. They are not constructed as a separate empirical population. Instead, they provide per-video observability descriptors that condition the interpretation of the RQ1 and RQ2 outputs and support the system-level design response formulated in Section 5.5.

The relevant variables include the ordinal `completeness_level`, the scalar `completeness_score`, the CDX-presence indicator `has_cdx`, the replayable-HTML indicator `has_html`, the WARC payload indicator `has_warc_payload` and the semantic-signal indicator `has_semantic_signal`. The completeness level, completeness score, `has_cdx`, `has_html` and `has_warc_payload` describe archival preservation evidence. The `has_semantic_signal` indicator describes whether a recoverable YouTube semantic signal is available for the same video identifier. Together, these variables describe archival preservation state and semantic observability. They do not encode platform intent, moderation rationale, content value, or preservation priority.

Table 5.3 summarises the analytical role of the datasets and feature matrices used in this chapter.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 5.3: Analytical datasets and their role in Chapter 5

Dataset or output	Primary use	Main constraint
RQ1 availability-trajectory feature matrix	Descriptive reconstruction of observable availability loss, archival presence, archival absence, and right-censoring	Supports observed patterns only; does not establish platform-side causes
RQ2 content attribute feature matrix	Logistic regression of final platform-observed semantic availability against content-level attributes	Supports conditional associations only; excludes causal interpretation
Observability artefact outputs	RQ3 design response and Design Cycle evaluation of system adequacy	Represents archival preservation state and semantic observability, not content value, deletion cause, or moderation rationale

5.3 RQ1: Quantitative Patterns of Availability Loss and Observability

This section addresses **RQ1**: *What externally observable patterns of deletion affect Syria-related videos on YouTube?*

The analysis is strictly descriptive and quantitative. It documents corpus-level regularities in how availability loss and disappearance are externally *observed* across platform unavailability and archival observability. No causal inference, intent attribution, or temporal survival modelling is performed. All results are derived exclusively from the frozen `rq1_feature_matrix`. In this section, deletion is used as a descriptive umbrella term for externally observable loss of platform availability, without implying any specific mechanism, intent, or moderation process.

5.3.1 RQ1 Analytical Population and Inclusion Criteria

This subsection defines the analytical population used for RQ1. The unit of analysis is the individual video identifier. Videos are included in the RQ1 analytical population when a platform-facing endpoint observation is available. Archival trajectory descriptors are added where CDX, archived HTML evidence and WARC payload evidence are available. Videos without archival history remain included in the RQ1 population, but are represented as structurally unobserved with respect to archival transitions. No filtering is applied based on content category, platform state, or outcome.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 5.4: RQ1 analytical population and archival observability subsets

Metric	Videos
Total videos (RQ1 analytical population)	50,022
Videos with archival evidence (CDX snapshots present)	18,170
Videos with observed archival availability signal ($t_{\text{avail}} \neq \emptyset$)	14,037
Videos with observed archival unavailability signal ($t_{\text{unavail}} \neq \emptyset$)	1,215
Videos with archival right censoring ($t_{\text{unavail}} = \emptyset$)	16,955

5.3.2 Platform-Unavailable Videos and Archival Coverage

This subsection reports platform-facing availability states observed through live YouTube endpoint checks and relates them to archival observability. Platform states are treated as externally observable outcomes at a single point in time. They are not interpreted as indicators of moderation intent.

Table 5.5: Distribution of YouTube platform availability states within the RQ1 analytical population

Endpoint availability label	**Videos**
available	22,987
error	22,451
age_restricted	2,175
unknown	1,321
account_terminated	1,088

Table 5.6 reports videos classified as semantically unavailable on YouTube and lacking Wayback history, grouped by their endpoint availability label. This cross-tabulation identifies cases where semantic unavailability in the platform-facing layer coincides with the absence of Wayback history, while the endpoint label records the specific response category observed by the live check.

Table 5.6: Video-level cross-tabulation of HTML-derived YouTube unavailability and absence of Wayback history, grouped by endpoint availability label

Endpoint availability label	**Videos**
error	10,051
age_restricted	1,765
available	1,542
account_terminated	908

5.3.3 Observed Archival Unavailability and No Observed Archival Unavailability

This subsection distinguishes between videos for which archival unavailability is directly observed and videos for which no archival unavailability is observed within the observational window. Observed unavailability corresponds to an explicitly detected explicitly unavailable state within archival evidence.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

All remaining cases are cases in which archival unavailability is not observed within the available evidence. They therefore remain unconfirmed with respect to archival unavailability.

Table 5.7: Observed versus unobserved unavailability within the RQ1 analytical population

Outcome	Videos
No archival unavailability observed	48,807
Observed archival unavailability	1,215

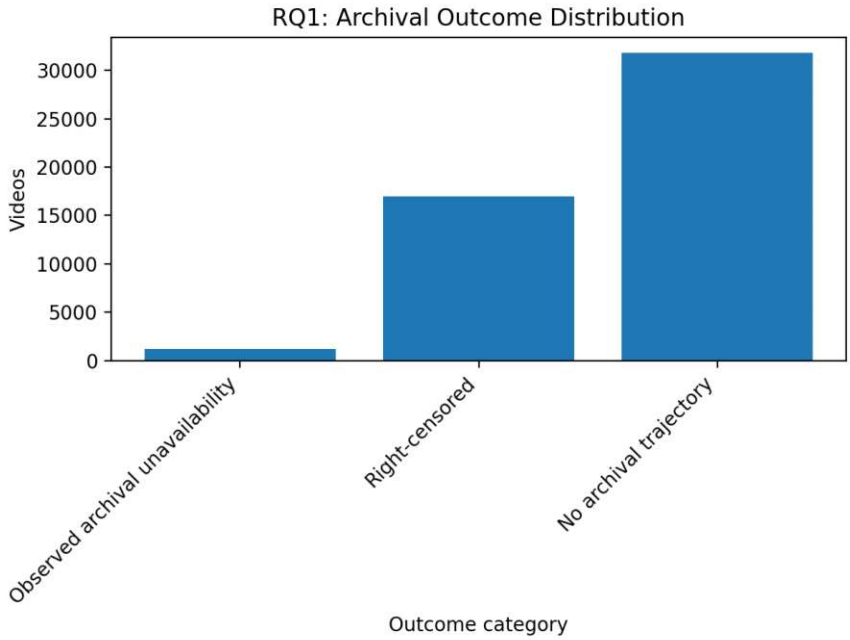

Figure 5.1: Distribution of observed archival unavailability and cases with no observed archival unavailability in the RQ1 analytical population.

5.3.4 Archival Absence as an Observability Pattern

This subsection documents the dominance of archival absence across the corpus. Archival absence is operationalised using snapshot count and archival lifespan indicators derived from CDX metadata. These measures describe limits of observability rather than content persistence.

Table 5.8: Archival presence and absence within the RQ1 analytical population

Condition	Videos
Videos with archival trajectory	18,170
Videos with no archival trajectory	31,852

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

5.3.5 Video-Level Evidence Membership Across Archive-Side Evidence States

This subsection reports video-level evidence membership across the archive-side evidence-state model. Membership is non-exclusive: a video may exhibit evidence for more than one state across its archival record.

Table 5.9: Non-exclusive video-level evidence membership across archive-side evidence states

Evidence state	Videos
Available	14,037
Restricted	5
Eroded	0
Unavailable	1,215
Redirect-only	579
Archival gap	2,833
Terminal absence	71
Unknown archive evidence	4,780

Unknown archive evidence records videos with at least one archived observation that cannot be assigned to one of the seven interpretable archive-side evidence states. This category is also non-exclusive. A video may therefore have unknown archive evidence and, at the same time, evidence for another archive-side state such as Available, Unavailable, Redirect-only, or Archival gap. The 4,780 videos reported as Unknown archive evidence are therefore not limited to videos with only unknown evidence.

The evidence-state counts in Table 5.9 are not expected to sum to the number of videos with archival history because membership is non-exclusive.

Table 5.10: Archival observability boundary conditions within the RQ1 analytical population

Boundary condition	Videos
Total RQ1 analytical population	50,022
No archival trajectory	31,852
Archived with at least one interpretable evidence-state membership	15,365
Archived without interpretable evidence-state membership	2,805

Videos with no archival trajectory have no CDX-based archival record and therefore fall outside the archive-side evidence-state model. Videos with at least one interpretable evidence-state membership provide evidence for one or more of the seven interpretable states.

The 2,805 videos listed as archived without interpretable evidence-state membership are videos for which archive-side evidence exists, but none of that evidence supports assignment to one of the seven interpretable archive-side evidence states. This group consists of 2,729 videos with unknown-only snapshots and 82 videos with a single redirect snapshot plus unknown snapshots.

The distinction between these 95 videos and the Redirect-only evidence state is important. A single redirect snapshot is not sufficient for Redirect-only membership. Redirect-only requires a repeated redirect

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

pattern, defined here as at least two consecutive redirect captures without an archived YouTube page available for inspection. The 95 single-redirect cases therefore remain in the boundary category because they contain archival evidence, but not enough evidence to support Redirect-only membership or any other interpretable archive-side evidence state.

5.3.6 Structural Fragmentation of Archival Traces

This subsection examines structural fragmentation among videos with archival presence. Fragmentation is assessed using counts of CDX temporal gaps exceeding one day between successive CDX snapshots. These CDX temporal gaps are distinct from the Archival gap evidence state reported in Table 5.9. These indicators describe discontinuities in archival traces without implying continuous monitoring, platform-side change, or confirmed content disappearance during the gap intervals.

Table 5.11: CDX temporal fragmentation among videos with Wayback history

Condition	Videos
No CDX temporal gaps (> 1 day)	13,574
At least one CDX temporal gap (> 1 day)	4,596

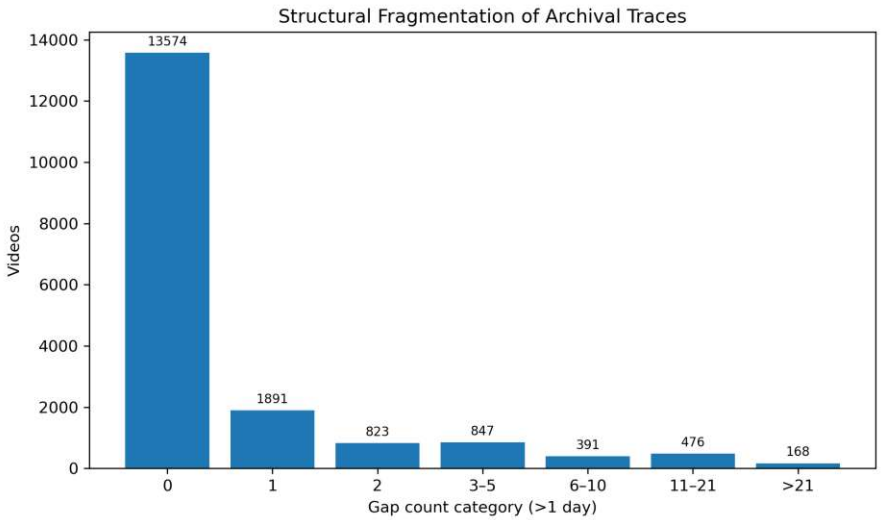

Figure 5.2: Binned distribution of CDX temporal gap counts among videos with Wayback history. Gap counts use the implemented one-day threshold and are grouped to show the zero-heavy distribution and the long right tail without excluding cases.

5.3.7 Summary of Measurable RQ1 Patterns

Taken together, the results document recurring observational patterns of partial observability across platform and archival layers. The RQ1 results distinguish between archive-side evidence-state membership, residual unknown archive evidence, no archival trajectory, and structural CDX temporal fragmentation. CDX temporal gap counts report one-day structural discontinuities and are not identical to the Archival gap evidence state. These patterns define the empirical conditions under which deletion can be observed, without making claims about platform intent or causal mechanisms.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.4 RQ2 Regression Analysis of Availability Outcomes

This section addresses **RQ2**: *How do factors such as topic, language, and metadata correlate with the likelihood of platform content removal on YouTube?*

This section examines statistical associations between final platform-observed availability outcomes on YouTube and content-level attributes using the frozen RQ2 content feature matrix. The analysis is strictly descriptive and evaluates whether observable differences in availability outcomes are statistically associated with pre-existing content characteristics. No causal inference, attribution of platform intent, or moderation rationale is implied.

5.4.1 RQ2 Modelling Population

The RQ2 content attribute feature matrix contains 49,635 videos with at least one admissible semantic-top assignment. For regression modelling, these videos are joined with the final HTML-derived YouTube availability outcome.

The outcome has three observed states: `playable`, `explicitly_unavailable` and `indeterminate`. The binary regression model uses only the two observable outcome states: `playable` and `explicitly_unavailable`. The 1,315 videos classified as `indeterminate` are excluded before model estimation. The final RQ2 regression population therefore contains 48,320 videos.

Table 5.12: Video-level construction of the RQ2 regression population from HTML-derived YouTube availability outcomes

Stage	**Videos**
Videos with admissible RQ2 content features	49,635
`playable` outcome	27,700
`explicitly_unavailable` outcome	20,620
Excluded: `indeterminate` outcome	1,315
Final RQ2 regression population	48,320

5.4.2 Model Specification

Regression models are specified with a binary dependent variable indicating final platform-observed semantic availability. Independent variables consist exclusively of content-level attributes derived from the frozen RQ2 feature matrix, including semantic-top category assignments. Explanatory variables are not derived from the observed availability outcome itself.

Let $Y_i \in \{0, 1\}$ denote the final platform-observed semantic availability status of video i, where $Y_i = 1$ indicates that the video is playable and $Y_i = 0$ indicates explicit unavailability at the end of the observation window, as determined by HTML-derived playability signals. Let $\mathbf{X}_i$ denote a vector of content-level attributes associated with video i.

Associations between content attributes and availability outcomes are modelled using a generalised linear model with a logit link function:

$$\log\left(\frac{\Pr(Y_i = 1 \mid \mathbf{X}_i)}{1 - \Pr(Y_i = 1 \mid \mathbf{X}_i)}\right) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}, \tag{5.1}$$

where β_0 denotes an intercept term and $\boldsymbol{\beta}$ is a vector of coefficients corresponding to the content attributes in $\mathbf{X}_i$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

All models are estimated using L1-regularised maximum likelihood to mitigate high dimensionality and collinearity arising from categorical feature expansion. Regularisation is used to automatically reduce the number of content categories in the model; the remaining coefficients are reported only as descriptive associations with final platform-observed availability outcomes. Robustness is assessed by estimating models under two regularisation strengths ($\lambda = 0.1$ and $\lambda = 0.3$). Unless stated otherwise, reported coefficients correspond to the more strongly regularised specification ($\lambda = 0.3$).

All model estimates are interpreted exclusively as conditional statistical associations and do not support causal claims regarding content moderation, enforcement mechanisms, or platform governance.

5.4.3 Results and Effect Sizes

This subsection reports descriptive associations between final platform-observed availability outcomes and content-level attributes. All results are derived from the frozen RQ2 feature matrix and are interpreted strictly under externally observable constraints.

Semantic attributes. This subsection examines associations between final platform-observed availability outcomes and semantic content attributes derived from the RQ2 taxonomy. Coefficient estimates are obtained from an L1-regularised logistic regression model under the more strongly regularised specification ($\lambda = 0.3$), with robustness assessed by re-estimating the model at $\lambda = 0.1$. Reference categories correspond to the modal semantic-top category per axis.

For clarity, semantic categories are grouped according to whether they are associated with higher unavailability (lower odds of playability) or lower unavailability (higher odds of playability). Unlike the geographic specification, no semantic category is eliminated under stronger penalisation; coefficients remain non-zero and directionally consistent across both regularisation levels. To support interpretability in the main text, Table 5.13 reports a reduced set of semantic attributes with the largest coefficient magnitudes under the $\lambda = 0.3$ specification. This reduction serves as a presentation choice and does not imply statistical significance or causal relevance.

Table 5.13: Selected semantic content attributes and their conditional associations with final platform-observed availability outcomes ($\lambda = 0.3$).

Axis	Semantic category	Coefficient	Odds ratio
Higher unavailability (lower odds of playability)			
Axis A	Human casualties and losses	-1.429	0.240
Axis A	Humanitarian activity	-0.635	0.530
Axis C	Political activity	-0.332	0.717
Axis C	Repression and violations	-0.242	0.785
Axis B	Martyrs and fatalities	-0.061	0.941
Axis C	Artistic and cultural expression	-0.037	0.964
Lower unavailability (higher odds of playability)			
Axis C	Non-event subject-referential content	+0.631	1.879
Axis A	Displacement and humanitarian crises	+0.306	1.358
Axis A	Protest and civic mobilisation	+0.259	1.296
Axis A	Armed violence and combat	+0.228	1.256

Several semantic categories are associated with reduced odds of playability; among these, *Human casualties and losses* has the largest negative coefficient among the semantic attributes. A small number of additional categories, including *Humanitarian activity*, follow at a substantial distance, while most other negative

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

associations are weak. On the positive side, *non-event subject-referential content* stands out as the strongest association with higher playability, with other positive effects remaining moderate.

Taken together, semantic attributes retain a consistent signal under regularisation, yet the overall pattern remains fragmented. A few categories stand out, but associations are distributed across multiple semantic dimensions and do not converge into a single, coherent content-based explanation of availability outcomes.

5.4.4 Robustness and Sensitivity Checks

Robustness is assessed by re-estimating regression models under alternative regularisation strengths. Coefficient stability across specifications is evaluated based on retention (non-zero coefficients) and sign consistency between the $\lambda = 0.1$ and $\lambda = 0.3$ models. Observed differences in coefficient magnitude across specifications are interpreted as model sensitivity rather than substantive change in underlying platform behaviour.

Exploratory specifications including geographic metadata yielded weak and uneven associations and are therefore not reported in the main analysis.

5.5 RQ3: System-Level Design Response Under Observability Limits

This section addresses **RQ3**: *Under external observability limits, which system-level design improvements can support digital memory preservation and alignment with Digital Humanism?*

The preceding analyses show that availability loss cannot be represented as a single binary event within the available evidence. Across platform-facing and archival layers, the observed record is partial, temporally fragmented, and shaped by limits of external observability. The archival completeness construct specified in Section 3.5.7 provides the system-level response used for **RQ3**: it represents archival evidence as layered, incomplete, and externally bounded.

5.5.1 Empirical Output of the Observability Artefact

This subsection documents the concrete empirical outputs produced by the system-level artefact introduced in response to **RQ3**. Rather than describing design intentions or architectural considerations, the subsection reports the observable variables, distributions, and structural properties generated by the artefact and made available for analysis. The purpose is to make the artefact explicitly documented within the thesis text and tables, independent of external code repositories or implementation details.

Unit of output and data freeze. Each video identifier corresponds to a single output record generated according to the predefined calculation rules. All values reported in this subsection are derived from the frozen feature dataset described in Section 4.9. No additional aggregation, filtering, or transformation is applied at the reporting stage beyond simple counting and proportion calculation.

Artefact Output Variables. For each video identifier v, the artefact produces the following observability-related variables:

- `completeness_level`: an ordinal categorical variable encoding the highest empirically observable preservation state.
- `completeness_score`: a continuous scalar in the interval $[0, 1]$ encoding degree of archival observability within that state.
- `has_cdx`: boolean indicator of CDX index presence.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

- `has_html`: boolean indicator of replayable HTML snapshot presence.
- `has_warc_payload`: boolean indicator of WARC-level payload presence.
- `has_semantic_signal`: boolean indicator of recoverable YouTube semantic signals (e.g., player response).

These variables are derived from externally observable evidence and describe archival preservation state and semantic observability. They do not encode platform intent, moderation rationale, content value, or preservation priority.

Distribution of Completeness Levels. Table 5.14 reports the empirical distribution of ordinal completeness levels across the analytical corpus.

Table 5.14: Distribution of ordinal archival completeness levels

Completeness level	Videos	Proportion
Full (CDX + HTML + WARC)	0	0.000
Partial (CDX + HTML)	17,926	0.358
Trace (CDX only)	244	0.005
Absent (no archival traces)	31,852	0.637
Total	50,022	1.000

The table shows that no video in the analytical corpus achieves full archival observability. A substantial proportion of the corpus is preserved only partially or at index level, while the majority lacks any archival traces. This confirms empirically that digital memory preservation operates under graded, non-binary conditions.

No video reaches the "Full" preservation level because CDX records in the analysed corpus do not expose the offset or filename information required to retrieve corresponding WARC payloads. Full payload-level preservation is therefore empirically defined but remains observationally inaccessible under the archival conditions encountered in this study.

Distribution of Completeness Scores. In addition to the ordinal completeness level, the artefact assigns a scalar completeness score that refines the representation of archival observability. The score captures variation in archival density and temporal span while remaining bounded by the ordinal preservation state defined in Section 3.5.7. Table 5.15 summarises the distribution of this score across the corpus.

Table 5.15: Summary statistics for completeness score

Measure	min	p25	median	p75	max
Completeness score	0.000	0.000	0.000	0.445	0.670

The score distribution is strongly right-skewed. Most observations are concentrated at zero, reflecting the large share of videos without observable archival traces. A subset of videos exhibits comparatively higher archival observability, but no case reaches the Full preservation level under the evidence available in this study. The scalar score therefore complements the ordinal completeness level by capturing variation in archival density and temporal span within the limits imposed by the categorical preservation state.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Archival Completeness and Semantic-Signal Availability. Table 5.16 cross-tabulates the archival completeness level with semantic-signal availability. The completeness level describes the archival preservation state of each video identifier. The semantic-signal indicator describes whether a recoverable platform-level semantic signal is available for the same video identifier.

Table 5.16: Cross-tabulation of archival completeness level and semantic-signal availability

Completeness level	Videos with semantic signals	Proportion within level
Full	0	0.000
Partial	15,016	0.838
Trace	2	0.008
Absent	0	0.000
Total	15,018	0.300

The table shows an uneven distribution of semantic-signal availability across archival completeness levels. Most videos with recoverable semantic signals are classified as Partial, where replayable archived HTML evidence is available. Two videos classified as Trace also contain recoverable semantic signals in the platform-observation layer. Across the full corpus, 15,018 videos contain a recoverable semantic signal, corresponding to 30.0% of all analysed videos. The cross-tabulation therefore documents how archival preservation state and semantic observability vary across the corpus.

Interpretive Status of the Artefact Output. The variables reported above are not treated as outcome measures, quality indicators, or proxies for content persistence. They encode conditions of observability under which availability-related observations are made. As such, they function as interpretive qualifiers that condition interpretation in subsequent statistical analyses rather than as explanatory or evaluative metrics.

By exposing graded preservation states and explicitly preserving incompleteness, the artefact operationalises a system-level response to **RQ3**: it improves the preservability of digital memory not by enforcing persistence, but by stabilising the traces through which memory can be empirically reconstructed under platform opacity.

5.5.2 Bounded Answer to RQ3

The system-level design response to **RQ3** is the construction of an observability-based preservation artefact that represents archival evidence as layered, incomplete, and externally bounded. Under external observability limits, the thesis identifies five system-level design improvements. First, availability evidence should be represented through layered observability states rather than through a single binary available/unavailable distinction. Second, partial traces should be retained as analytical outcomes instead of being removed through completeness-based filtering. Third, missingness and indeterminacy should propagate explicitly through the feature construction process rather than being imputed as availability or loss. Fourth, all transformations from raw traces to analytical variables should be reproducible and auditable through fixed rules and run-level provenance records. Fifth, availability-related outputs should remain non-causal and non-normative, so that they document the conditions under which platform-mediated digital memory can be observed without attributing intent, responsibility, or content value.

This answer is bounded by the evidence available to the observational measurement system. It does not prescribe platform moderation policy, evaluate individual removal decisions, or claim to recover content that is not observable in the available archival and platform-facing sources. The system-level contribution consists in making observability limits explicit and analytically usable, so that empirical claims about platform-mediated digital memory remain traceable to the evidence from which they are derived.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.6 Evaluation of the Observational Measurement System

This section evaluates the observational measurement system as the methodological artefact that supports all three research questions. The evaluation is not limited to the completeness artefact formulated for **RQ3**. Instead, it assesses whether the implemented system produces valid, reproducible, and bounded outputs for: (i) the descriptive reconstruction of observable availability and loss patterns required for **RQ1**; (ii) the construction of feature matrices and statistical modelling population required for **RQ2**; and (iii) the derivation of system-level design implications required for **RQ3**.

The evaluation follows the Design Cycle evaluation structure defined in Section 3.6. Analytical evaluation, functional testing, simulation-style corpus application, and descriptive informed argument provide the evaluation methods. Criteria C1–C5 provide the standards used to judge artefact adequacy. The descriptive informed argument is operationalised through structured criteria mapping, in which evidence from Chapters 4 and 5 is mapped to C1–C5.

5.6.1 Evaluation Scope and Object

The evaluated object is the observational measurement system as an implemented Design Science Research instantiation. The evaluation also includes the supporting artefacts through which the system produces the analytical outputs used in this thesis: the archive-side evidence-state model, the trajectory reconstruction procedure, the feature extraction procedure, the graded completeness representation, and the reproducibility instrumentation. The evaluation therefore concerns the adequacy of the measurement system for producing bounded empirical and design-relevant knowledge under platform opacity. It does not evaluate YouTube governance, moderation policy, archival institutional performance, or the normative importance of individual videos.

Table 5.17 summarises the objects included in the evaluation and identifies which research-question outputs each object supports.

Table 5.17: Objects evaluated in the Design Cycle evaluation reported in Chapter 5

Evaluated object	**Role in the thesis**	**RQ support**
Observational measurement system	Implemented instantiation that collects, stores, derives, and exposes externally observable platform and archival evidence	RQ1, RQ2, RQ3
Archive-side evidence-state representation	Construct and method for representing observed availability evidence, unavailability evidence, CDX temporal gaps, and right-censoring	RQ1
Content attribute feature matrix	Analytical bridge between human taxonomy features and platform-observed availability outcomes	RQ2
Completeness artefact	Graded representation of archival observability and partial preservation states	RQ3
PRIMAD and validation outputs	Reproducibility and transparency support for data derivation and analysis	RQ1, RQ2, RQ3

5.6.2 Evaluation Inputs and Evidence Base

The evaluation uses evidence produced by the implemented measurement system, evidence from the evaluation checks reported in this section, and empirical outputs reported in Chapters 4 and 5. The

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

relevant inputs are: the frozen feature dataset from Section 4.9; the RQ1 availability-trajectory feature matrix; the RQ2 content attribute feature matrix; the RQ3 completeness outputs; the reported descriptive counts, regression coefficients, and observability distributions; and the reproducibility artefacts described in Chapter 3. No independently verified ground-truth dataset exists for item-level YouTube availability across the full observation window, and the evaluation therefore does not use external correctness labels. Instead, it evaluates internal consistency, reproducibility, preservation of partial evidence, and compatibility between the system outputs and the bounded claims made for each research question.

The analytical-evaluation component checks the internal consistency of the artefact specification before the criteria mapping is applied. It concerns the formal archive-side evidence-state structure, the distinction between CDX-level and HTML-level evidence, the trajectory reconstruction rules, the separation of observable states from causal claims, and the completeness bounding rule. These checks establish whether the system specification is internally coherent and whether its outputs can be interpreted within the external-observation boundary defined in Chapter 3.

The functional-testing component of the evaluation checks whether the implemented measurement system produces reproducible and contract-valid outputs from fixed inputs. Table 5.18 summarises the functional tests corresponding to the evaluation procedure defined in Section 3.6.

Table 5.18: Functional-testing evidence for the observational measurement system

Functional test	Evidence used	Result
Independent re-run	Fixed inputs were reprocessed on the defined validation subset to verify deterministic output reproduction.	Fulfilled
Schema validation	Derived datasets and feature matrices were checked against predefined data contracts.	Fulfilled
PRIMAD manifest checks	Run manifests record provenance, instrument versions, method parameters, data sources, seeds, and checksums.	Fulfilled
Feature-matrix freeze	RQ1 and RQ2 analyses use frozen analytical matrices without ad hoc post-processing at the analysis stage.	Fulfilled

The simulation-style evaluation is implemented through application of the measurement system to the fixed analytical corpus. The resulting aggregate outputs are not compared against an external ground-truth label set, because no such externally verified item-level ground truth exists for the observation window. Instead, the evaluation checks whether the full-corpus outputs exhibit the expected artefact properties: differentiated observability states, preserved partial traces, explicit right-censoring and missingness, reproducible feature construction, and non-causal interpretability.

5.6.3 Support for RQ1

For **RQ1**, the relevant evaluation question is whether the system validly supports the descriptive reconstruction of observable availability-loss and observability patterns. The RQ1 results show that the system produces per-video availability and archival-observability outputs for the full analytical population of 50,022 video identifiers. It distinguishes platform-facing availability states, archival presence, archival absence, observed archival unavailability, and right-censoring. It also preserves videos without archival history as structurally unobserved with respect to archival transitions rather than excluding them or treating them as confirmed disappearances.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The system therefore supports the RQ1 answer because it reconstructs measurable patterns without converting unobservable cases into deletion claims. The relevant limitation is that the system can report only observed availability and archival states; it cannot establish the complete deletion history of videos or identify the cause of unavailability. This limitation is not treated as a failure of the system, because the thesis explicitly restricts RQ1 to externally observable patterns. Table 5.19 summarises the evaluation of system support for RQ1.

Table 5.19: Evaluation of system support for RQ1

Requirement for RQ1	**Evidence from the system outputs**	**Assessment**
Reconstruct observable availability patterns	Platform availability states are produced for the full RQ1 analytical population, while archival trajectory variables and observability subsets are populated where Wayback evidence is available	Fulfilled
Represent partial and indeterminate states	Right-censored videos and videos without archival history are retained as explicit analytical states	Fulfilled
Avoid causal overclaiming	RQ1 results are reported descriptively and do not infer moderation intent or platform-side mechanism	Fulfilled
Maintain reproducible derivation	Outputs are derived from the frozen feature dataset and predefined trajectory variables	Fulfilled

5.6.4 Support for RQ2

For **RQ2**, the relevant evaluation question is whether the system validly supports bounded statistical association analysis between final platform-observed availability outcomes and content-level attributes. The RQ2 analysis defines a clear unit of analysis and separates content-feature availability from outcome observability. Videos without admissible semantic-top assignments are excluded during feature-matrix construction, and videos with `indeterminate` HTML-derived outcomes are excluded before model estimation. The dependent variable is derived from HTML-level playability signals. The regression model uses content-level attributes that are constructed before modelling and are not derived from the observed availability outcome itself. The resulting coefficients are interpreted as conditional associations rather than as causal explanations.

The system therefore supports the RQ2 answer because it provides a reproducible feature matrix, an explicit modelling population, and a bounded interpretation of model outputs. The main limitation is that the regression analysis cannot identify moderation mechanisms, policy causes, or hidden platform decisions. This limitation is handled by the model specification and interpretation rules, which restrict RQ2 to observable associations within the analysed population. Table 5.20 summarises the evaluation of system support for RQ2.

5.6.5 Support for RQ3

For **RQ3**, the relevant evaluation question is whether the system supports a system-level design response to the observability limits identified through RQ1, RQ2, and Chapter 4. The completeness artefact

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 5.20: Evaluation of system support for RQ2

Requirement for RQ2	**Evidence from the system outputs**	**Assessment**
Define a reproducible modelling population	RQ2 inclusion and exclusion counts are specified through semantic-top feature availability rules and HTML-derived outcome filtering	Fulfilled
Construct outcome and predictors from observable data	The dependent variable is derived from final platform-observed playability signals; predictors are content-level taxonomy attributes	Fulfilled
Prevent causal interpretation	Model results are explicitly interpreted as conditional statistical associations only	Fulfilled
Expose feature and population constraints	Semantic-top exclusions, indeterminate outcome exclusions, and robustness checks are reported within the RQ2 section	Fulfilled

represents archival observability through ordinal levels and a bounded scalar score. It preserves CDX-only traces, replayable HTML traces, semantic-signal availability, and absence of archival traces as distinct conditions. It also specifies non-use constraints that prevent the artefact from being interpreted as a proxy for content value, deletion, suppression, or platform moderation.

The system therefore supports the RQ3 answer because it translates observed measurement limits into a concrete representation that can be used to condition interpretation and preserve partial traces. The main limitation is that the artefact does not improve platform-side transparency or recover unavailable content. Its contribution is limited to the observational measurement infrastructure developed in this thesis. Table 5.21 summarises the evaluation of system support for RQ3.

Table 5.21: Evaluation of system support for RQ3

Requirement for RQ3	**Evidence from the system outputs**	**Assessment**
Derive design response from observed limits	The artefact responds to archival absence, partial HTML replayability, semantic-signal loss, and cross-source observability limits	Fulfilled
Represent graded preservation states	Completeness levels distinguish full, partial, trace, and absent archival observability states	Fulfilled
Preserve non-use constraints	Completeness is not used as a proxy for content value, platform intent, or deletion cause	Fulfilled
Remain externally observable and non-causal	Inputs are restricted to CDX, HTML, WARC-retrievability, and replay-layer evidence	Fulfilled

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.6.6 Structured Criteria Mapping Against C1–C5

This subsection operationalises the descriptive informed argument through structured criteria mapping. Evidence from the analytical evaluation, functional testing, simulation-style corpus application, and empirical results is mapped to criteria C1–C5. The purpose of this mapping is to make explicit whether the observational measurement system satisfies the requirements defined in Section 3.6 across **RQ1**, **RQ2**, and **RQ3**. The mapping therefore evaluates the whole measurement system, not only the **RQ3** completeness artefact.

Table 5.22: Structured criteria mapping of the observational measurement system against C1–C5

Criterion	RQ1 support	RQ2 support	RQ3 support	Status
C1 Observability differentiation	Distinguishes platform states, archival presence, observed unavailability, gaps, and right-censoring	Prevents interpretation of the dependent variable as a complete deletion record	Grounds layered representation of CDX, HTML, WARC, and semantic-signal evidence	Fulfilled
C2 Preservation of partial evidence	Retains videos with CDX temporal gaps, right-censoring, and no archival history	Documents feature exclusions and preserves modelling-population boundaries	Preserves partial traces as explicit completeness levels and scores	Fulfilled
C3 Analytical operability without causal attribution	Reports measurable patterns only	Reports conditional associations only	Derives design implications without inferring platform intent or moderation cause	Fulfilled
C4 Reproducibility and transparency	Uses rule-based trajectory and feature derivation	Uses a frozen feature matrix and specified model configuration	Uses fixed completeness rules and reproducibility instrumentation	Fulfilled
C5 Alignment with problem framing	Studies content availability under external observability limits	Bounds statistical interpretation under missingness and platform opacity	Converts observed limits into a system-level design response without normative scoring	Fulfilled

Across all five criteria, the structured criteria mapping supports the judgement that the observational measurement system satisfies the functional requirements defined in Section 3.6. This judgement is based on evidence from the analytical specification, functional-testing results, simulation-style corpus application, and empirical outputs reported for **RQ1**, **RQ2**, and **RQ3**. The system differentiates observability states, preserves partial evidence, supports quantitative analysis without causal attribution, produces reproducible and transparent outputs, and remains aligned with the problem framing established by the research gaps.

5.6.7 Evaluation Result, Discrepancy Log, and Remaining Limits

The Design Cycle evaluation shows that the observational measurement system adequately supports the three research questions within the limits of external evidence. This judgement is based on the full evaluation structure defined in Section 3.6: analytical evaluation, functional testing, simulation-style corpus application, and descriptive informed argument operationalised through structured criteria mapping. It supports **RQ1** by producing reproducible descriptive patterns of platform availability,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

archive-side evidence-state membership, residual unknown archive evidence, observed unavailability, right-censoring, and archival absence. It supports **RQ2** by producing bounded feature matrices and association models that can be interpreted without inferring platform-side causes. It supports **RQ3** by translating observed limits of external evidence into a system-level design response that preserves partial traces and makes incompleteness explicit.

This evaluation does not establish platform-internal ground truth. It evaluates whether the artefact system can produce reproducible, analytically usable, and non-causal representations of platform-mediated digital memory under conditions of platform and archival opacity. The remaining limits are therefore reported as boundaries of the artefact's claim rather than as implementation failures.

Table 5.23 records the discrepancies and unresolved boundaries identified during the Design Cycle evaluation. These entries define the conditions under which the observational measurement system can produce valid claims.

Table 5.23: Discrepancy log and remaining evaluation boundaries

Discrepancy or boundary	Evaluation consequence	Treatment
No external item-level ground truth for platform availability over the observation window	Correctness cannot be evaluated against independently verified deletion or moderation labels.	Bounded claim
No access to platform-internal moderation records	The system cannot attribute unavailability to moderation intent, policy enforcement, legal action, uploader action, or technical failure.	Non-causal interpretation
Incomplete and uneven archival coverage	Absence of archival evidence cannot be interpreted as absence of the original video or as confirmed disappearance.	Explicit missingness and right-censoring
Partial semantic yield in archived HTML	Replayable HTML does not always expose video-level semantic signals.	Separate CDX, HTML, and semantic-signal indicators
No completed field deployment with the constituent community	Practical utility for preservation organisations remains unvalidated in an operational setting.	Future Relevance Cycle closure

Across these boundaries, the evaluation shows that the system makes platform-mediated digital memory more observable and analytically tractable without removing the opacity that defines the research problem. Its adequacy is therefore limited to external observation: the system differentiates observable states, preserves partial and indeterminate evidence, supports bounded statistical analysis, and records reproducibility conditions, but it does not recover unavailable content, explain platform decisions, or validate practical use in preservation workflows.

5.7 Summary of Findings and Evaluation

This chapter has reported the statistical outputs derived from the frozen feature datasets, the **RQ3** observability artefact outputs, and the Design Cycle evaluation of the observational measurement system. The chapter therefore connects the empirical and statistical answers to **RQ1** and **RQ2** with the system-level design response and artefact-system evaluation required by the Design Science Research framework.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

For **RQ1**, the analysis documented measurable patterns of partial observability across platform and archival layers, including platform-facing unavailability, archival absence, observed archival unavailability, right-censoring, and structural fragmentation of archival traces. These results show what can be observed externally, without claiming complete deletion histories or platform-side causes.

For **RQ2**, the regression analysis identified conditional associations between final platform-observed semantic availability and selected content-level attributes. The strongest reported semantic association with lower odds of playability is the category *Human casualties and losses*, while other associations remain weaker and distributed across multiple semantic dimensions. These results are interpreted as bounded statistical associations within the observed modelling population, not as causal evidence of moderation or suppression.

For **RQ3**, the chapter formulated a system-level design response based on graded archival observability. The completeness artefact represents CDX presence, replayable HTML, WARC-retrievability, semantic-signal availability, and archival absence as structured conditions of observation. Its purpose is to preserve and expose the limits of digital-memory observability, not to rank content or infer platform decisions.

The Design Cycle evaluation reported in Section 5.6 shows that the observational measurement system supports all three research questions against criteria C1–C5. The system differentiates observability states, preserves partial evidence, remains analytically operable without causal attribution, supports reproducibility and transparency, and aligns with the problem framing of external observation under platform opacity.

The evaluation provides three outputs specified in Section 3.6: an evaluation matrix mapping C1–C5 to evidence across **RQ1**–**RQ3**, system-level design implications derived from the evaluated measurement system, and a discrepancy log documenting unresolved boundaries. The evaluation confirms the adequacy of the observational measurement system for producing bounded, reproducible, and non-causal outputs under platform and archival opacity. It does not validate platform-internal causes, complete archival preservation, or field-level operational utility.

Substantive interpretation of these findings, including their implications for digital memory and platform-mediated forgetting, is developed in Chapter 6.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 6

Discussion

6.1 Purpose and Scope of the Discussion

This chapter brings together the empirical results reported in Chapters 4 and 5 and explains what these results show for the empirical study of platform-mediated digital memory.

Without introducing new empirical claims beyond the reported results, the chapter has articulated how the observed results define the limits of what can be established through external observation under conditions of platform and archival mediation.

All statements in this chapter are based solely on (i) the observational design and measurement constraints defined in Chapter 3 and (ii) the quantitative and descriptive results reported in Chapters 4 and 5.

6.2 Consolidated Empirical Findings on Observability Limits

This section restates the main empirical results reported in the preceding chapters. These results are treated as direct outputs of the measurement process rather than as theoretical or interpretive claims. For clarity, they are presented as three consolidated empirical findings (E1–E3), each directly linked to the corresponding measures, tables, and figures in Chapters 4 and 5.

6.2.1 E1: Persistent Empirically Unconfirmed Availability States Despite Use of All Available Observation Sources

Chapters 4 and 5 show that most videos with archival history remain right-censored at the end of the observation window. Within the CDX-defined archival universe ($n = 18{,}170$), a terminal unavailable state is directly observed for only 1,215 videos (6.7%), while 16,955 videos (93.3%) remain right-censored (Table 4.10; Figure 4.9; Table 5.1).

Right-censoring is used here in the operational sense defined in Section 4.6: a descriptor is classified as right-censored when its final observed state does not indicate unavailability, meaning that disappearance cannot be confirmed using the available observations within the observation window. This right-censored condition is distinct from terminal absence. Terminal absence requires evidence that the replayable

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

archived HTML record ends while later archive-side observations remain present. Right-censoring instead records that no terminal unavailability condition is observed within the available observation window. Accordingly, this empirically unconfirmed state is recorded as a direct measurement outcome rather than being treated as missing, incomplete, or failed data. The statistical results reported in Chapter 5 preserve the same distinction between videos with directly observed unavailability and videos for which disappearance cannot be empirically confirmed (Table 5.7).

6.2.2 E2: Infrastructure-Mediated Absence of Observable Signals Without Attribution to Human Decisions

Results reported in Chapter 4 show consistent non-alignment between CDX records, archived HTML snapshots, and live platform responses. These non-alignments can be accounted for by directly observable behaviour of archival infrastructure and replay systems and do not require any assumptions about moderation intent or human decision-making.

First, the presence of a CDX record and the availability of a replayable HTML snapshot represent two separate observation layers. Across the corpus, the Wayback CDX index contains 54,013 snapshot records, while only 28,602 replayable HTML snapshots were retrieved at the same timestamps. As a result, 25,411 CDX snapshot records (47.0%) do not have a corresponding replayable HTML capture at the same timestamp (Table 4.11; Table 4.3).

Second, observation at the live platform layer is limited by measurement constraints. A total of 1,321 live platform checks were classified as `unknown` because they consistently returned HTTP status code 429 with the error type `rate_limited`. In these cases, the returned responses consist of interstitial pages that do not expose video availability information (Section 4.5.2; Table 4.9). These cases are therefore treated as platform states that cannot be measured using the available observations, rather than as evidence of availability, restriction, or removal.

Third, archived HTML replay produces technically valid pages across multiple response types, including archive-generated interstitial pages and non-target responses that replace platform content at the time of observation. Within the archived HTML corpus, 1,062 snapshots (3.7%) consist of Wayback rate-limit notices, and additional snapshots are classified as robots exclusion pages or captured site error pages (Table 4.5). [1]

In addition, the archive-side evidence-state model records residual unknown archive evidence where archival evidence exists but cannot be assigned to one of the seven interpretable archive-side evidence states. This residual state accounts for 29,378 snapshot-level observations. Of these, 18,888 have no matching parsed HTML result, while 10,490 have a parsed HTML result but still do not provide sufficient evidence for assignment to an interpretable state. At video level, 2,805 archived videos remain assigned only to residual or insufficient evidence. These cases are therefore not treated as missing data or as availability states, but as observable limits of archive-side interpretation.

Taken together, these results show that non-alignment across archival and platform observation layers can arise from the behaviour of archival and replay infrastructure itself. These infrastructure-related limits define what can and cannot be determined using external observation alone (Section 4.7).

The structural sources of non-alignment documented in E2 share a common property: they arise from automated systems — whether Wayback replay mechanisms, platform rate-limiting responses, or CDX indexing procedures — that operate without producing externally visible records of their decision logic. This property is relevant to the Vienna Manifesto's Principle 5, which holds that decisions with

[1] Wayback rate-limit notices, robots exclusion pages, and captured error pages are archive-generated or archive-mediated responses that replace the original platform content during replay, meaning that the archive infrastructure itself can prevent observers from accessing platform-specific meaning even when an archived record exists.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

consequences for individual or collective human rights must continue to be made by humans rather than delegated to automated systems [WLA+19]. The thesis does not evaluate whether any specific automated system has operated appropriately or inappropriately. What E2 documents, within the strictly observational scope of this study, is that the cumulative effect of automated systems operating across multiple layers of the platform–archive infrastructure is a pattern of signal absence that cannot be attributed to any specific decision or mechanism from external evidence alone. The structural opacity that results from this cumulative effect is the empirical condition that the Manifesto's Principle 5 addresses at a normative level and that this thesis addresses at a measurement level.

6.2.3 E3: Loss of Video-Level Semantic Information Within Archival Replay Systems

Chapters 4 and 5 provide direct empirical evidence that loss of video-level semantic information can occur within the archival replay process itself, even when technically valid HTML pages are available. Across the archived HTML snapshots ($n = 28{,}602$), 28,567 snapshots (99.88%) contain complete canonical HTML elements, indicating that the pages are technically well formed (Table 4.4). However, platform-specific semantic information is not consistently present: only 17,757 snapshots (62.1%) expose a detectable YouTube player response, while 10,845 snapshots (37.9%) do not expose any player response (Table 4.6).

In addition, video-context metadata is often missing. Only 16,241 snapshots (56.8%) contain all listed metadata fields, while 39.9% lack a channel identifier and 43.2% lack an owner channel name (Table 4.7). At the video level, cross-source comparison further shows that a substantial subset of videos has no archived HTML snapshot containing video-level semantic information prior to the live platform check, with the proportion differing across live platform states (Table 4.14).

The RQ3 observability artefact reported in Chapter 5 makes this uneven loss of semantic information explicit at the corpus level. In the ordinal completeness classification, no videos fall into the *Full* preservation level (CDX + HTML + WARC), while 31,852 videos (63.7%) are classified as *Absent* and 17,926 videos (35.8%) as *Partial* (Table 5.14). Within the same artefact outputs, recoverable semantic signals are mainly present in the *Partial* level and are rare in the *Trace* (CDX-only) level (Table 5.16).

Taken together, these results show that the presence of replayable archival material does not ensure that video-level meaning can be recovered. Loss of semantic information is an empirically observable property of the measurement and replay systems used to document platform content, rather than a property of platform behaviour alone.

6.3 Relationship to Existing Conceptual Frameworks

This section relates the empirical findings E1–E3 to the conceptual frameworks introduced in Chapter 2. Its purpose is to state which aspects of these frameworks are addressed by the findings and which are not. No judgement is made about the correctness, validity, or normative adequacy of the frameworks. Accordingly, the analysis below is limited to comparing E1–E3 with the explicit problem definitions and conceptual emphases stated in the cited texts.

To ensure comparability across frameworks, each subsection follows the same rubric: (1) framework emphasis (as relevant to this study), (2) mapping of E1–E3 to that emphasis, (3) boundary conditions (what the findings cannot address), and (4) the net contribution within that framing. Table 6.1 provides an at-a-glance mapping of E1–E3 to each framework; the subsections specify emphasis, boundary conditions, and contribution.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 6.1: E1–E3 mapped to framework emphases (short clauses).

Framework	E1: Right-censoring	E2: Layer misalignment	E3: Semantic loss
Connection level	transparency limit	inspectability discrepancy	loss of inspectable signals
Digital Humanism	Non-confirmable disappearance under external access.	Externally visible discrepancies across layers.	Observable loss of content-level signals in replay.
Connection level	temporal reconstruction limit	mediation-based reconstruction limit	semantic reconstruction limit
Infosphere	Finite-window limit on historical confirmation.	Layered mediation constrains reconstruction.	Missing semantic/context signals in valid records.
Connection level	apparatus-dependent observability limit	network-distributed infrastructural effect	intermediation-induced content transformation
Posthuman	Observational apparatus and relational subject formation shape what disappearance events can be confirmed [Bar07, Bra13].	Heterogeneous network components each contribute independently to signal absence [Lat05].	Technical intermediation transforms content across crawl, storage, and replay layers [Hay17].

6.3.1 Digital Humanism

Framework emphasis. Digital Humanism presents an explicit diagnosis of contemporary digitalisation and formulates this diagnosis as a call for increased accountability and corrective action [WLA+19, WPLG22, Wer23]. Within this thesis, Digital Humanism is used to delimit the class of problems under consideration to those that concern transparency- and accountability-relevant properties of digital systems when studied through external observation. The empirical analysis itself remains limited to what can be observed from platform-facing requests and external archive interfaces when studying platform-mediated digital memory.

Mapping of E1–E3. Table 6.1 summarises how E1–E3 address this emphasis, based on the empirical results reported in Sections 6.2.1–6.2.3.

Not addressed / boundary. The findings are limited to externally observable traces and cannot establish claims that require access to platform-internal or archive-internal states. Accordingly, enforcement decisions, policy-triggered internal availability states, and archive-internal ingestion or replay transformations that are not externally visible cannot be supported within the observational design of this study (Chapter 3; Sections 4.7 and 4.6).

Net contribution within this framing. Within the Digital Humanism research motivational foundation established in Chapter 2, the thesis contributes a quantitative description of observability limits in platform-mediated memory research; a measurement framework that operationalises transparency requirements at the level of research infrastructure; and a reproducible artefact that enables the non-evaluative study of automated system effects on documentary accessibility [WPLG22, Wer23].

6.3.2 Infosphere and Hyperhistory

Framework emphasis. Floridi describes contemporary digital technologies as forces that form an informational environment (the “infosphere”) and argues that modern societies increasingly rely on digital

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

information and information infrastructures for their normal operation [Flo14]. Within this account, the recording, processing, and transmission of information are directly linked to the conditions under which history and memory can be constructed and reconstructed in digitally mediated contexts [Flo14]. Floridi further emphasises that digital records require semantic interpretation and that loss, inaccessibility, or opacity of information infrastructures constrains what can be known and reconstructed [Flo14].

Mapping of E1–E3. Table 6.1 summarises how E1–E3 instantiate these reconstruction constraints in the platform–archive setting analysed here, based on Sections 6.2.1–6.2.3.

Not addressed / boundary. The findings are limited to externally observable traces and cannot establish claims that require access to platform-internal or archive-internal states. Accordingly, semantic meaning, provenance, and internal processing histories cannot be established beyond what is externally observable in archived representations and replay outputs; the results remain a description of constraints on reconstruction under partial access.

Net contribution within this framing. In this respect, the thesis provides a platform-specific case of how layered and partially opaque information infrastructures constrain what can be confirmed or reconstructed from digital records [Flo14].

6.3.3 Posthuman

Framework emphasis. Posthuman scholarship addresses socio-technical systems through a concern for distributed agency, material entanglement, and the role of non-human components in shaping outcomes [Bra13, Lat05, Bar07, Hay17]. Braidotti and Latour argue that outcomes are emergent properties of heterogeneous networks of human and non-human actors rather than effects attributable to any single cause or decision point [Bra13, Lat05]. Barad further holds that the observational apparatus is constitutively involved in producing what it detects, while Hayles argues that content is actively transformed rather than neutrally transmitted across technical layers [Bar07, Hay17].

Mapping of E1–E3. Table 6.1 summarises how E1–E3 can be interpreted within this framework 6.2.1–6.2.3.

Not addressed / boundary. The findings are limited to externally observable traces and cannot establish claims that require access to platform-internal or archive-internal states. Accordingly, actor-level intent, internal decision points, and governance actions that are not externally observable through the measurement stack cannot be recovered from the data artefacts used in this study; the analysis remains at the level of distributed effects across the coupled infrastructure rather than internal causal accounts.

Net contribution within this framing. In this respect, the thesis provides a platform-specific empirical characterisation of how observability is shaped by the coupled and heterogeneous operation of the platform–archive measurement stack, in which no single component determines the observable record [Lat05, Bar07, Hay17].

6.3.4 Cross-framework synthesis

The framework comparison conducted in Sections 6.3.1 through 6.3.3 reveals a pattern of convergence that is analytically significant beyond the individual mappings. Each of the three frameworks — Digital Humanism, the infosphere account, and the posthuman — addresses the empirical findings E1 through E3 through a different conceptual vocabulary, yet each arrives at a structurally similar conclusion about the conditions under which digital memory can be empirically studied.

Digital Humanism situates E1 through E3 within a concern for transparency and accountability of digital systems: the right-censoring, infrastructure-mediated signal absence, and semantic loss documented in this thesis reflect the conditions produced when digital systems operate without making their decision logic externally visible. The appropriate research response, from this perspective, is to design measurement frameworks that make these conditions of opacity explicit and analytically tractable [WLA+19].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The infosphere account situates E1 through E3 within a concern for the conditions under which historical reconstruction is possible: if digital records are fragmented, semantically degraded, and structurally incomplete, then what can be reconstructed from them is constrained not by the significance of the original events but by the infrastructural conditions under which they were captured and preserved [Flo14]. The appropriate research response, from this perspective, is to make those infrastructural constraints explicit and to bound analytical claims accordingly.

The posthuman perspective situates E1 through E3 within a concern for distributed agency and socio-material entanglement: the observability limits documented in this thesis are produced not by a single actor or system but by the coupled operation of multiple technical and institutional components — crawl schedules, replay mechanisms, moderation algorithms, rate-limiting responses — each of which contributes to shaping the observable record without requiring attribution to any individual human decision [Lat05, Bar07]. The appropriate research response, from this perspective, is to characterise the distributed effects of the infrastructure without attributing them to a single cause or agent.

These three methodological conclusions are not identical but they are compatible, and their compatibility demonstrates something important about the empirical findings: E1 through E3 are not merely statistical results about a particular corpus but characterisations of conditions that any study of platform-mediated digital memory must confront, regardless of the theoretical framework through which it approaches the problem. The convergence of three theoretically distinct frameworks on the same methodological conclusion is evidence that those conclusions are robust rather than framework-dependent. This thesis contributes not only to the three individual literatures but to their integration: by demonstrating that empirical findings about observability limits are interpretable within all three frameworks, it establishes that those frameworks address a common empirical problem from different directions, and that the problem is substantive enough to warrant coordinated attention across the disciplines they represent.

The frameworks differ in what they foreground when situating these constraints. Digital Humanism foregrounds transparency- and accountability-relevant properties of digital systems under restricted access; the infosphere and hyperhistory account foregrounds how informational environments and semantic interpretability condition what can be reconstructed from records; and the posthuman perspective foregrounds mediation and distributed effects across socio-technical assemblages. These differences do not alter the empirical scope of E1–E3, but they provide distinct conceptual languages for situating the same externally measured limits of observability.

6.4 Observability Conditions Across Frameworks

Taken together, the results restate three empirically observed conditions that remain as measurement-grounded constraints within the scope of this study: (i) most videos remain right-censored even when all available observation sources in the observational measurement system are applied (Section 6.2.1); (ii) limits of observability arise from the behaviour of infrastructural and replay layers, which shape what can be observed through external measurement (Section 6.2.2); and (iii) loss of video-level semantic information occurs within archival replay and measurement systems as part of the observation process itself (Section 6.2.3).

These conditions describe how observation operates in this study. They are reported as direct measurement outcomes grounded in the analyses presented in Chapters 4 and 5. They are presented as empirical findings that characterise the observation process and its outcomes within the defined analytical scope.

6.5 Implications for Empirical Research on Digital Memory

The findings imply that empirical research on platform-mediated digital memory must explicitly account for limits of observability as substantive empirical outcomes. First, right-censoring and empirically uncon-

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

firmed disappearance should be preserved and reported as explicit measurement outcomes rather than being treated as missingness or measurement failure (Section 6.2.1). Second, cross-source non-alignment should be treated as an infrastructural and replay-layer condition of observation, requiring explicit differentiation between index-level traces, replayable representations, and semantically interpretable signals (Section 6.2.2). Third, the semantic completeness and contextual completeness of archived representations should be measured and reported independently of structural validity, since structurally valid archived HTML frequently omits video-level semantic signals and contextual metadata (Section 6.2.3).

The three methodological implications identified in this section acquire additional significance when situated against the specific empirical domain examined in this thesis. The observability limits documented across archival and platform-facing layers apply to a corpus of Syria-related YouTube videos spanning 2011 through 2025 — a corpus that represents a substantial portion of the audiovisual documentation of one of the most widely observed and extensively platform-documented armed conflicts of the early twenty-first century [Ris19, Whi24, Tar24].

The finding that 93.3 per cent of videos with archival history remain right-censored means that for the overwhelming majority of the corpus, it is not possible to determine from external evidence whether the documentation they contain remains accessible to those who would seek to use it for historical research, legal investigation, or collective memory formation. The unavailability that cannot be confirmed is not merely a statistical residual: it is the empirical form taken by digital amnesia as introduced in Chapter 2 — a collective-level reduction in the accessible digital record that cannot be confirmed in its full extent without access to platform-internal data, but whose conditions are empirically observable. The finding that 63.7 per cent of the analytical population have no archival traces, and that 43.2 per cent of archived snapshots lack an owner channel name, further characterises the semantic degradation through which archival representations lose the contextual information needed to interpret, verify, or use their content [DH18, Ris19].

The regression analysis of final platform-observed availability outcomes reported in Chapter 5 adds a further dimension to these implications. The conditional associations identified between content attributes and availability outcomes — notably the substantial negative association between the content category of human casualties and losses and the odds of being observed as available — are reported as statistical patterns within the limits of observability established in Chapter 4. They do not establish causation, do not identify mechanisms, and do not support inferences about specific moderation decisions. Their methodological implication is nonetheless significant: they suggest that distributional patterns of platform-observed availability are not uniform across content categories, and that the study of differential availability requires analytical frameworks capable of representing both the statistical associations and the observability limits that constrain their interpretation. The Vienna Manifesto's Principle 2 identifies the safeguarding of information dissemination as a central concern of platform governance [WLA+19]. Within the observational scope of this thesis, the finding that availability patterns differ across content categories is relevant to this concern not as a normative verdict but as a measurement result that establishes what empirical research can and cannot determine about differential accessibility under conditions of platform opacity.

The empirical finding that platform-mediated changes in content availability operate through automated systems whose decisions leave no externally observable trace is directly relevant to the Vienna Manifesto's Principle 5, which holds that decisions with consequences for individual or collective human rights must continue to be made by humans and that automated decision-making systems should support rather than replace human judgment [WLA+19]. This thesis does not assess whether any specific automated decision was appropriate or advocate for any particular governance response. What it documents, within the constraints of external observation, is that the absence of externally observable decision traces is a structural property of the measurement environment that cannot be resolved by more sophisticated analytical methods operating on the same external signals. The implication for empirical research on digital memory in high-stakes documentation contexts is that measurement frameworks must be designed to make this structural property explicit and analytically tractable rather than treating it as noise or missing data.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

At the corpus level, the RQ3 artefact outputs further imply that observability conditions can be made consistently representable across the dataset by explicitly encoding graded preservation states and by preserving incompleteness rather than collapsing heterogeneous traces into binary indicators (Section 5.5.1). Within the observational design of this thesis, these implications concern methodological practice and measurement representation rather than platform-side explanation or intervention.

6.6 Limitations

Limits of observation arising from the measurement setup. The study is subject to several limits that arise from how observation is technically performed. First, the observation window is finite, which fixes the time period within which availability trajectories can be reconstructed. Second, all observations rely on platform-facing interfaces and web archive systems rather than direct access to platform-internal data. Third, historical access depends on archival indices and replay systems that determine which representations are retrievable at the time of observation. Together, these conditions shape which signals can be observed and how availability, absence, and archival presence are recorded. The empirical effects produced by these conditions—specifically right-censoring, cross-source non-alignment, and graded semantic loss in archived representations—are measured and reported as results in Sections 6.2.1–6.2.3.

A further limitation concerns the residual state unknown archive evidence. This state is not an error category, but an explicit archive-side observability boundary. It records cases where archival evidence exists, but the available observations do not support assignment to one of the seven interpretable archive-side evidence states. In the analysed corpus, 4,780 videos contain at least one unknown archive observation, and 2,805 archived videos have no interpretable evidence-state membership. These cases show that the presence of archival records does not necessarily imply that archive-side evidence is semantically interpretable.

Limits of corpus construction arising from data access and selection. The study is also shaped by limits related to how the analytical corpus is constructed. First, the corpus is defined by data access provided through a non-governmental organisation, over which the researcher has no influence with respect to scope, selection criteria, or completeness. Second, only a subset of the organisation's stored material is publicly accessible and available for empirical analysis. Third, access to archival records is governed by institutional policies that determine whether payload-level artefacts can be inspected. As a result, the dataset reflects the boundaries of third-party data provision rather than exhaustive coverage of the underlying content universe. These conditions affect corpus coverage and representativeness and are treated as explicit properties of the study design.

6.7 Summary

This chapter has synthesised the empirical results presented in Chapters 4 and 5 and clarified their significance for the study of platform-mediated digital memory. Without introducing new data or interpretation, the chapter has articulated how the observed results define the limits of what can be established through external observation under conditions of platform and archival mediation. By relating the empirical findings to existing conceptual frameworks only within a clearly defined analytical scope, the chapter has positioned the results as measurement-grounded evidence about the conditions under which digital memory can be observed, documented, and compared. This positioning is consistent with the Digital Humanism research foundation established in Chapter 2, which motivates the study of platform-mediated memory through the lens of transparency and accountability rather than normative evaluation [WLA+19]. This synthesis provides the basis for the concluding chapter.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 7

Conclusion

7.1 Summary of Findings

This thesis examined the persistence, disappearance, and observability of Syria-related YouTube videos using externally observable platform and archival data collected between 2011 and 2025. The study addressed three research questions through a rule-based observational measurement system and quantitative analysis.

With respect to **RQ1**, the analysis showed that most videos with archival history do not reach an empirically observable end state of unavailability within the observation window. Only a small fraction of videos exhibit an explicitly observed unavailable state in archival records, while the majority remain right-censored. This result shows that disappearance frequently cannot be confirmed through external observation alone, even when multiple data sources are used.

With respect to **RQ2**, statistical analysis identified conditional associations between final platform-observed availability outcomes and content-level attributes derived from human-assigned taxonomy labels. These associations are unevenly distributed across semantic categories and remain fragmented. They do not converge into a single explanatory pattern and are conditioned by limits of observability documented at the platform and archival layers.

With respect to **RQ3**, the thesis implemented and evaluated a system-level artefact that represents graded conditions of archival observability. The artefact differentiates between index-level presence, replayable representations, and semantic signal availability, and encodes these conditions using ordinal completeness levels and a bounded scalar score. The empirical outputs show that full archival observability is not achieved for any video in the corpus and that partial and absent preservation states dominate.

7.2 Contributions

This thesis makes three bounded contributions.

Empirical contribution. The study provides a large-scale empirical account of platform-mediated digital memory based on externally observable evidence. Using a corpus of over fifty thousand videos, it documents the prevalence of right-censoring, cross-source non-alignment, and semantic loss within archival replay. These observations document conditions under which digital content cannot be fully verified or reconstructed through public interfaces.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Methodological contribution. The thesis introduces an observability-based measurement approach that treats absence, indeterminacy, and partial archival traces as explicit empirical outcomes. By distinguishing between archival index records, replayable HTML representations, and recoverable semantic signals, the approach avoids collapsing heterogeneous evidence into binary indicators. The completeness artefact operationalises these distinctions in a form that is compatible with quantitative analysis and reproducible across datasets.

In this respect, the methodological contribution extends beyond its technical scope to enact, at the level of the research infrastructure, the transparency and accountability orientation that the Vienna Manifesto identifies as a structural requirement for socially responsible technological systems: by making the derivation conditions of analytical measures fully explicit and independently verifiable, the framework ensures that the limits of what the measurement process can show are visible to — and critically assessable by — the communities whose practice the research is designed to inform [WLA+19].

Practical contribution. The thesis contributes a reproducible observational measurement system that embeds transparency into the measurement process itself rather than implementing it as a supplementary measure. The system records evidence provenance explicitly, produces auditable analytical datasets whose conditions of construction are independently verifiable, and preserves graded visibility states rather than imputing completeness. The system-level design implications derived from observed preservation gaps and infrastructural constraints, together with the design principles and evaluation criteria documented in the study, function as meta-artefacts that can be adopted and evaluated by future researchers working in analogous observational contexts. Field testing of the measurement system by its constituent community — digital memory scholars, web archiving researchers, and civil society organisations engaged in the long-term preservation of conflict-related documentation — constitutes the primary direction for future Relevance Cycle closure.

Taken together, these contributions align with the Digital Humanism orientation of the thesis by making the conditions of platform-mediated digital memory observable, reproducible, and explicitly bounded. The empirical contribution documents what can be observed from external platform and archival traces; the methodological contribution specifies how such observations can be produced without collapsing partial evidence into binary claims; and the practical contribution provides a reproducible measurement system that can be evaluated and adapted in future preservation and research contexts.

7.3 Future Research

This thesis shows that content availability evidence and archival completeness can be partially reconstructed from external traces. It also shows the limits of that approach. External traces make it possible to observe evidence of unavailability, partial archival records, residual unknown archive evidence, and loss of semantic content in replayed pages. However, these traces do not provide a complete account of why such changes occur. They also do not guarantee that the resulting observations can be used directly in documentation practice, regulatory assessment, or legal settings.

The results of this thesis therefore do not justify a simple extension of the same measurement task. Future research should not focus only on collecting more of the same observations over a longer time period. That would improve coverage, but it would not address the main open problems identified here. These problems arise once external observability has already been established but remains incomplete. They concern five separate questions. First, it remains unclear how the measurement system performs in real preservation and documentation workflows. Second, the present analysis is retrospective and does not show whether future loss can be anticipated. Third, observed disappearance still cannot be linked reliably to formal platform actions. Fourth, the thesis does not define how archival traces should be interpreted when they are used as evidence. Fifth, the whole method still depends on fragmented archival infrastructures that were not built for systematic and reproducible comparative analysis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

For these reasons, the next stage of research should move from measurement alone to use, prediction, attribution, interpretation, and infrastructure. The five directions below follow directly from the limits identified in the empirical analysis and are formalised in the project concepts listed in Appendix C.

7.3.1 Operational Deployment and Field Validation

The system developed in this thesis was evaluated under controlled research conditions. The thesis shows that availability trajectories, completeness levels, and observability patterns can be derived in a reproducible way from external traces. However, this does not show that the system is already useful in real documentation workflows. The results remain research outputs unless they can be applied by organisations that collect, preserve, or assess digital records in practice [UU20].

This limitation matters because the practical value of the system depends on how its outputs are understood and used. In the thesis, completeness levels distinguish between records that are fully recoverable, partly recoverable, or missing from the available archives. In a real operational setting, these distinctions would not remain descriptive only. They would influence decisions about what should be preserved first, which cases require manual review, and where gaps in the archival record are serious enough to affect later interpretation. The present work does not test whether such decisions can actually be taken on the basis of the system output alone or whether users require additional explanation and contextual knowledge.

Future research should therefore test the system in collaboration with organisations that work with digital records under real constraints [MMK22]. This includes archives, documentation groups, and organisations that monitor public-interest content [AK20]. The main task is to examine whether the outputs produced by the system can be used directly in practice without creating new forms of error. This requires observation of actual use, not only technical evaluation. It must include how users read the outputs, how these outputs fit into existing workflows, where misinterpretation occurs, and which parts of the system require additional explanation or redesign.

Such work would move beyond the question of whether the method works in principle. It would address whether the method remains reliable once it enters institutional practice. This is especially important because a system that is technically reproducible can still fail if its outputs are interpreted too strongly or used outside the limits of the evidence. The next step is therefore not only deployment, but field validation under practical conditions. This direction is developed in Appendix C.1 as the project TRACE-DEPLOY.

7.3.2 Prediction of Availability Loss

The empirical results of this thesis are retrospective. They reconstruct availability and archival completeness after changes have already occurred. This is appropriate for the research questions of the thesis, but it leaves an important practical problem unresolved. If content loss is only documented after the fact, then the method remains limited to description. It does not support early intervention.

Several results in this thesis indicate that availability loss is not a single and instantaneous event. Instead, it often appears as a process that unfolds over time. Temporal gaps, delayed archival decay, divergence between sources, and declining semantic quality in replayed pages all suggest that loss may have observable precursors [Pew24, KVdSS+14]. These patterns are visible in the data analysed here, but the thesis does not test whether they can be used predictively. It therefore remains unknown whether external traces can support not only retrospective reconstruction but also prospective risk assessment.

Future research should examine whether changes in archival traces can be used as early indicators of later disappearance [Zub18, HKS+15]. This work must remain careful and modest in its claims. The goal is not to promise certain prediction. The goal is to test whether measurable changes in capture frequency, replay quality, or cross-source consistency provide useful risk signals before a video or page becomes unavailable [STvO+18]. This requires longitudinal modelling, explicit treatment of censoring, and systematic comparison between cases that later disappear and cases that remain available.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

This direction would add a new layer to the present method. The thesis shows how to reconstruct what can still be observed. Future work should test whether the same external traces can also be used to estimate future instability. If successful, such work would be relevant not only for research but also for preservation practice, because it would help identify content that should be prioritised before further loss occurs. This direction is developed in Appendix C.2 as the project TRACE-PREDICT.

7.3.3 Regulatory Attribution and Evidentiary Linkage

A central finding of this thesis is that externally observed disappearance is not self-explanatory. When content becomes unavailable, the available traces do not show whether the cause was platform moderation, user deletion, account loss, archival failure, replay degradation, or another process outside direct observation. This means that the thesis can describe disappearance, but not assign it to a specific formal action.

This limitation becomes more important in the current European regulatory environment. The DSA Transparency Database provides formal records of moderation actions, including Statements of Reasons and related metadata for content decisions [Eur26]. These records create a new possibility that was not available in earlier platform governance contexts: they allow external researchers to compare observed availability changes with formally recorded platform actions. However, this possibility is not yet a solution. The presence of both data sources does not by itself establish a valid connection between them.

Future research should therefore develop explicit rules for linking external observability traces to regulatory records. The key problem is not only technical matching. It is evidentiary discipline. A link between an observed disappearance and a moderation record must be justified in a way that does not overstate causal claims. This requires clear criteria for when two records count as a plausible match, how strong such a match is, and how uncertainty should be documented. Some cases may support strong linkage. Others may support only weak linkage, or none at all. The method must make these differences explicit.

This line of work would not remove the broader problem of partial observability. However, it would reduce one of its most important consequences. At present, disappearance is observable but not attributable. Regulatory linkage would create a more disciplined basis for supervisory analysis and for research on platform governance under conditions of incomplete evidence. This direction is developed in Appendix C.3 as the project TRACE-LINK. It is closely related to earlier work on platform opacity and the limited external visibility of content governance [SWQY19, KL20, GBK20, TFC25].

7.3.4 Evidentiary Interpretation and Standards

The thesis introduces graded completeness in order to describe how much archival material is available for a given record. This construct is analytically useful because it distinguishes between different levels of archival recovery. However, the thesis does not define how these levels should be interpreted when archived material is used as evidence in documentation, regulatory, or legal contexts.

This is a serious open problem because the evidentiary value of archived digital records is not self-evident. A record may be present in an archive and still remain incomplete, degraded, or difficult to interpret. A replayable page may preserve some information but omit other elements that are necessary for later assessment. A partial record may support some claims while leaving others unresolved. The present thesis uses completeness and observability as analytical variables, but it does not translate them into explicit rules for evidentiary use.

Future research should therefore define how archival traces can be used credibly in contexts where claims about provenance, integrity, and interpretation matter. This includes at least three tasks. First, it must specify which metadata are minimally required in order to describe an archived digital record in a way that supports later use. Second, it must define how provenance and transformation history should be documented when material has passed through multiple archival and processing steps. Third, it must

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

establish how uncertainty should be reported when records are incomplete, partly degraded, or missing elements that would otherwise support stronger claims.

This work would not replace legal or institutional judgement. Instead, it would clarify the evidentiary limits of archived traces and make their conditions of use explicit. That is necessary because archived platform material is increasingly used in documentation and open-source investigation, while its quality depends on heterogeneous traces whose meaning is often assumed rather than defined [Whi24, Mak21, Ris19]. This direction is developed in Appendix C.4 as the project TRACE-EVIDENCE.

7.3.5 Federated Observability Infrastructure

The method developed in this thesis depends on archival infrastructures that were not designed as a unified research environment. In practice, different archives capture different parts of the web, apply different collection priorities, offer different access interfaces, and preserve different levels of replayability. The Wayback Machine CDX index plays a central role in the present work, but it is only one part of the wider archival landscape [VdSNS13].

This creates both methodological and practical limits. Results may depend on which archive is queried, which replay system is available, and which interface exposes the relevant traces. Archives also differ in their technical and institutional conditions, which makes systematic comparison difficult [Rol14]. The thesis documents these limitations, especially where records are fragmented across sources or where replay remains technically possible but semantically reduced. However, it does not solve the infrastructure problem itself.

Future research should therefore build a federated observability infrastructure that combines multiple archival sources into a single analysis environment [JKVdS+21]. This requires more than aggregating data. It requires standardised access rules, explicit provenance recording, and a reproducible processing chain that makes clear how results were generated [HGA16, KSBVdS19]. The point is not to erase differences between archives. The point is to make them visible and analysable within one system.

Such an infrastructure would improve the reproducibility of observability-based research and reduce dependence on a single archival source. It would also make it easier to compare how different archives capture the same content and where their records diverge. This would strengthen both research reliability and later practical use. The problem here is infrastructural rather than analytical, and it therefore requires a distinct research and development effort. This direction is developed in Appendix C.5 as the project TRACE-INFRA.

Taken together, these five directions define the next stage of research that follows from this thesis. They do not repeat the same empirical task. Instead, they address the consequences of the thesis's main finding: digital memory on platforms can be studied empirically through external traces, but only within clear limits of observability. Future research must therefore move beyond measurement alone and address how such observations are used, predicted, linked, interpreted, and supported by the infrastructures on which they depend.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX A

Overview of Generative AI Tools Used

Table A.1: Overview of generative AI tools used

Tool	**Product name / version / date**	**Where used**	**How used**
ChatGPT	ChatGPT, GPT-5.5 Thinking, accessed on 1 June 2026	Selected thesis-related work, including language revision, structural consistency checks, LaTeX troubleshooting, and formulation support.	Used selectively as an auxiliary aid. The research design, data processing, analysis, interpretation, argumentation, and submitted wording were developed, checked, and finalised by the author.

No whole passages of text generated by generative AI tools were adopted without substantial changes. Therefore, no separate prompt documentation is required for unchanged AI-generated passages.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX B

Methodological Supplement

B.1 Literature grounding outputs

This section provides the full outputs of the literature grounding process described in Section 3.4.2 by listing the complete construct dictionary, variable list, and methodological precedents used in the thesis.

Table B.1: Construct dictionary derived from literature grounding

Construct	Definition in this thesis	Literature domain
Availability descriptor	Per-video record summarising externally observable availability information derived from archival and platform-facing evidence	Platform governance, web archiving
Semantic availability state	Availability classification derived from HTML-level playability signals	Web archiving, platform studies
Platform unavailability	Platform-facing condition in which a video is not accessible at the time of observation	Platform governance
Archival observability	Condition under which availability can be externally observed through archival records	Web archiving
Archival trajectory	Existence of at least one archival interaction record for a given video identifier	Web archiving
Metadata	Structured descriptive information about a digital object	Archival research, IS methodology
Metadata completeness	Extent to which essential descriptive metadata fields are present	IS measurement theory
Platform governance regimes	Rules and enforcement practices shaping content visibility	Platform governance, policy

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Construct	Definition in this thesis	Literature domain
Collective memory	Socially constructed record of shared experiences	Memory studies
Digital amnesia	Reduction in the accessible digital record over time	Digital memory studies
Digital Humanism	Framework aligning digital systems with human values and accountability	Digital Humanism
Temporal gaps / discontinuities	Gaps between archival observations exceeding defined thresholds	Web archiving
Content taxonomy features	Derived features used for classification and modelling	Methodology
Human-assigned taxonomy labels	Externally assigned labels used as analytical inputs	Methodology

Table B.2: Variable list derived from the observational pipeline

Variable / output	Definition in this thesis
t_{check}	Timestamp of platform availability check
$avail_v^{\text{endpoint}}$	Endpoint-derived availability classification
$avail_v^{\text{html}}$	HTML-derived availability classification
$S(v)$	Set of archival snapshots (t_i, $status_i$, $mime_i$)
$html_state_i$	Semantic availability state derived from archived HTML
$t_{\text{avail_start}}$	First observed available state
$t_{\text{restr_start}}$	First observed restricted state
$t_{\text{erode_start}}$	First observed erosion state
$t_{\text{unavail_start}}$	First observed unavailable state
t_{disp}	Analytical disappearance time
δ_v	Right-censoring indicator
$\{d_v^{(P)}\}_{P \in Q}$	Duration of each evidence-state interval
$\mathcal{C}(v)$	Complete availability descriptor
T_v	Time to observed unavailability
Archival lifespan	Time between first and last archival snapshot
CDX temporal gap count	Number of intervals greater than one day between consecutive CDX snapshots
Content taxonomy features	Regression-ready features derived from labels
Human-assigned taxonomy labels	External classification labels

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table B.3: Methodological precedents supporting the literature grounding

Function	**Representative sources**	**Zotero branch**
Review typology	Snyder (2019)	M1
Scoping and design	Kitchenham et al., systematic mapping literature	M2
Search and acquisition context	CDX API, Wayback, YouTube API, DSA database	M4
Interpretive review logic	Boell and Cecez-Kecmanovic (2014)	M5
Validation, design science, reproducibility	Hevner (2004, 2007), Wirth (2000), Straub (2004), Ferro et al., survival analysis literature	M6

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX C

Future Research Project Concepts

This appendix presents five project concepts derived from the main open problems identified in the thesis. These projects do not repeat the core measurement task. They begin from the assumption that external observability has already been established, but remains incomplete. Each project therefore addresses a different next-stage problem: use in practice, prediction, attribution, evidentiary interpretation, or infrastructure.

Each project is written as a separate research programme of sufficient scope for a multi-person team. The projects are distinct from one another in problem, method, and output. They are linked directly to the Future Research section in Chapter 7.

C.1 TRACE-DEPLOY: Operational Deployment of Observability Systems

Problem. The thesis demonstrates that content availability and archival completeness can be reconstructed from external traces in a reproducible way. However, the thesis does not test whether these outputs are usable in real documentation or preservation workflows. It therefore remains unknown whether the system can support actual decisions without additional interpretation or whether its outputs are too technical, ambiguous, or incomplete for direct operational use.

Objective. This project examines how the observability system performs when used by organisations that work with digital records in practice. The main goal is to determine whether the system outputs can be applied directly in real workflows, where they fail, and what kind of additional support is needed to use them responsibly.

Approach. The project would deploy the system in selected partner organisations that collect, preserve, or assess digital records. It would study how the system is integrated into existing workflows, how users interpret its outputs, and how system-generated assessments influence later decisions. The work would combine technical deployment, workflow analysis, and systematic observation of failure cases. It would also compare expected use with actual use in order to identify where the system remains too complex, too weak, or too easily misread.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Scope. The project focuses on operational use. It does not attempt to solve regulatory attribution, evidentiary standardisation, or archive federation as its main task. These issues may appear in practice, but they are addressed as separate projects elsewhere in this appendix.

Outputs. The project would produce a validated deployment model, workflow integration guidelines, a structured account of failure modes, and recommendations for the responsible practical use of observability outputs.

C.2 TRACE-PREDICT: Prediction of Availability Loss

Problem. The thesis reconstructs availability retrospectively. It shows that content loss often appears through temporal gaps, fragmentation, and other changes that become visible only after the fact. However, it does not test whether these observed changes can be used to estimate future risk.

Objective. This project examines whether availability loss can be anticipated from earlier changes in archival traces. The main goal is to determine whether measurable signals in capture history, replay quality, or cross-source divergence can serve as early indicators of later disappearance.

Approach. The project would analyse longitudinal archival records and model how observable signals change over time before content becomes unavailable. It would compare cases that later disappear with cases that remain available, while treating censoring and incomplete observation as part of the research design rather than as noise to be ignored. The method would include feature construction from archival traces, temporal signal analysis, and evaluation of predictive performance under realistic conditions of missingness.

Scope. The project focuses on risk estimation before loss occurs. It does not seek to explain why loss occurred in a regulatory or causal sense, and it does not define evidentiary standards for later use. Those questions belong to other projects in this appendix.

Outputs. The project would produce tested early-warning indicators, predictive models for availability loss, and evaluation results showing the limits and reliability of such prediction under partial observability.

C.3 TRACE-LINK: Regulatory Attribution and Linkage

Problem. A major limitation of the thesis is that observed disappearance cannot be assigned to specific platform-side actions from external traces alone. Although the DSA Transparency Database now provides formal records of moderation decisions, the existence of such records does not automatically solve the attribution problem.

Objective. This project develops methods for linking external observability traces to regulatory transparency records. Its main goal is to determine when an observed disappearance can be connected to a documented moderation action and how strong such a connection is.

Approach. The project would combine archival observations with regulatory records such as Statements of Reasons made available through the DSA Research API [Eur26]. It would define matching rules, test alternative linkage strategies, and classify the resulting links by strength and uncertainty. A central task would be to distinguish between cases where attribution is strongly supported, cases where only weak linkage is justified, and cases where no responsible link can be made. The project would therefore treat uncertainty as a core output rather than a residual issue.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Scope. The project focuses on attribution to formal platform actions. It does not define evidentiary standards for legal use, and it does not attempt to build a full federated archive infrastructure. Its main task is to move from observation to disciplined linkage.

Outputs. The project would produce a linkage method, explicit uncertainty rules, prototype tools for attributed trace analysis, and guidance for regulatory and research use under conditions of incomplete evidence.

C.4 TRACE-EVIDENCE: Evidentiary Interpretation of Archived Records

Problem. The thesis uses completeness and observability as analytical variables, but it does not define how archived digital traces should be interpreted when they are used as evidence. At present, the presence of a record in an archive is often treated as if its meaning were obvious, even though archived material may be incomplete, transformed, or difficult to interpret.

Objective. This project defines how archived traces can be described and evaluated when they are used in documentation, regulatory, or legal contexts. Its main goal is to establish what metadata, provenance information, and uncertainty documentation are required for credible evidentiary use.

Approach. The project would compare the requirements of different institutional settings in which archived digital records are used. It would examine how provenance, integrity, completeness, and semantic loss affect the claims that can be made about a record. On this basis, it would define minimum metadata requirements, a clear model for documenting transformation history, and explicit rules for reporting uncertainty. It would also test these requirements through expert validation and prototype implementation in selected use cases.

Scope. The project focuses on evidentiary meaning. It does not seek to predict future disappearance or attribute disappearance to platform actions. It also does not treat infrastructure federation as its primary problem, although provenance recording and system traceability remain relevant to its design.

Outputs. The project would produce metadata and provenance guidelines, uncertainty-reporting rules, validation examples, and a practical standard for the responsible evidentiary use of archived digital records.

C.5 TRACE-INFRA: Federated Observability Infrastructure

Problem. The thesis depends on archival infrastructures that are fragmented and uneven in coverage, interface design, and replay behaviour. This makes systematic comparison difficult and limits reproducibility across datasets, users, and institutional settings.

Objective. This project builds an integrated observability environment across multiple archival sources. Its main goal is to make observability-based analysis more reproducible, less dependent on a single archive, and more transparent in its processing chain.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Approach. The project would combine several archival sources into one technical environment while preserving explicit information about source differences. It would standardise data access, define common processing interfaces, and record how results are produced at each step. The work would include source integration, provenance-aware processing, and reproducibility support so that the same input can be analysed in a consistent and auditable way across different runs and users.

Scope. The project focuses on technical and infrastructural conditions of analysis. It does not test field use as its main task, and it does not define legal or regulatory standards. Those questions depend on the infrastructure, but are conceptually separate from building it.

Outputs. The project would produce a federated observability system, standardised access interfaces, reproducibility records for analysis runs, and a technical basis for later multi-archive research and practice.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX D

Derivation and Literature Positioning of the Archive-Side Evidence-State Model

D.1 Aim and scope

This appendix explains how the seven interpretable archive-side evidence states plus one residual unknown archive evidence state used in this thesis were positioned relative to prior literature and how they were derived from the evidence available to the study. The appendix has four aims. First, it reviews existing classification models that are relevant to online content availability, platform restriction, content persistence, and web-archive evidence. Second, it explains why existing models are not sufficient for the specific measurement problem of this thesis. Third, it defines the evidence model from which the seven interpretable archive-side evidence states plus one residual unknown archive evidence state are derived. Fourth, it explains the limited novelty claim of the *7+1 archive-side evidence-state model*.

The model contains seven interpretable archive-side evidence states:

1. Available,
2. Restricted,
3. Eroded,
4. Unavailable,
5. Redirect-only,
6. Archival gap,
7. Terminal absence.

The model also retains one residual state: Unknown archive evidence. This state is used when CDX or processed snapshot evidence exists, but the evidence is insufficient to assign the observation to one of the seven interpretable archive-side evidence states.

These states are used as archive-side evidence states. They are not used as causal explanations of YouTube moderation, removal, or restriction decisions. The 7+1 archive-side evidence-state model does not claim

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

that there are only seven real-world ways in which a YouTube video can become unavailable or difficult to access. It claims only that the seven interpretable archive-side evidence states plus one residual unknown archive evidence state are distinguished by the archive-side measurement model used in this thesis.

This appendix follows literature-review guidance that a review should map prior work, classify existing approaches, identify limitations, and make clear how a new model or contribution is grounded in the reviewed literature [Sny19, BCK14]. The purpose is therefore not to prove that no similar model exists anywhere. The purpose is to show that, within the reviewed literature, no prior model was identified that defines the same seven interpretable archive-side evidence states plus one residual unknown archive evidence state with the same evidence rules.

Boundary condition of the 7+1 archive-side evidence-state model. The *7+1 archive-side evidence-state model* applies only once an archive-side trajectory exists. If no CDX record exists for a video identifier, the archive provides no trajectory to classify and no archive-side evidence state is assigned.

D.2 Plain definitions of technical terms

This section defines the technical terms used in the appendix. The definitions are given before the state-of-the-art discussion because the appendix addresses readers who may not be familiar with web archiving or platform measurement.

Table D.1 makes two distinctions that are central for this thesis. First, a CDX record is not the same as a replayable archived HTML page. A CDX record shows that the archive recorded an interaction with a URL. A replayable archived HTML page means that saved page content can actually be opened and inspected. Second, a platform-facing signal is not the same as an archive-side signal. A platform-facing signal describes the current response from YouTube. An archive-side signal describes evidence preserved by the Internet Archive.

D.3 State of the art of existing classification models

Before defining a new model, it is necessary to check whether existing classification models already solve the same problem. The relevant literature is distributed across several areas. Platform-side models describe current platform states such as public, private, restricted, removed, or terminated. Persistence studies describe whether previously collected online items can still be retrieved later. Web-archive studies describe archive coverage, redirects, replay quality, missing resources, and retrieval limits. Digital memory and digital forgetting literature explains why loss, restriction, and incomplete access matter for memory-relevant digital records.

Table D.2 summarises these literature groups and their relevance to the *7+1 archive-side evidence-state model*.

As shown in Table D.2, the state of the art provides relevant parts of the classification problem, but these parts are distributed across different research areas. Existing literature does not provide one directly reusable model that combines platform restriction, archived page availability, metadata degradation, archive redirects, capture gaps, and terminal absence into a single archive-side classification model for YouTube pages.

D.4 Platform-side classification models

Platform-side classification models describe the current status of content on a platform. They are relevant because they show that online availability is not limited to a simple available/unavailable distinction.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table D.1: Basic terms used in the appendix

Term	Plain definition
Video identifier	The unique string that identifies a YouTube video. In a standard YouTube URL, it is the value after `watch?v=`.
Signal	One observable piece of evidence. Examples are an HTTP status code, an archive timestamp, a saved HTML page, a redirect record, or a platform-facing error message.
State	A label assigned by the measurement system after reading one or more signals. A state describes what the evidence shows. It does not explain why the platform behaved in that way.
CDX record	An Internet Archive index entry showing that the archive recorded a URL at a specific time. A CDX record may include a timestamp, HTTP status code, and file type. It does not by itself show the saved page content.
Replayable archived HTML page	A saved HTML page that can be opened and inspected through the archive. This is stronger evidence than a CDX record because the page content can be read and classified.
Platform-facing signal	A current response from YouTube at the time of measurement. It may indicate availability, restriction, account-level problems, or measurement failure.
Archive-side signal	Evidence from the Internet Archive, such as CDX records, redirect records, archived HTML pages, capture gaps, or the absence of later saved pages.
Over-time archive record	The ordered list of archive observations for one video, sorted by capture date.

Platforms may remove content, restrict access, reduce visibility, apply labels, or enforce account-level measures [Gil18, Gil22, GBK20].

Official YouTube documentation distinguishes several platform-facing states. The YouTube Data API includes fields related to privacy status, upload status, processing status, embeddability, and regional restrictions [Goo26]. YouTube Help documentation also describes age-restricted content, videos locked as private, video privacy settings, and channel or account termination [You26a, You26d, You26b, You26c]. These categories are relevant because they show that platform access can be restricted without being identical to deletion.

The Digital Services Act also distinguishes several types of platform action. It refers to removal, disabling access, restriction of visibility, demotion, suspension, and termination as distinct forms of platform response [Eur22]. The DSA Transparency Database further operationalises platform statements of reasons for such actions [Eur].

Table D.3 summarises the relevance of these platform-side classifications.

Table D.3 shows that platform-side models provide strong support for separating restricted access from unavailability. Their limitation for this thesis is that they classify current platform states, policy categories, or regulatory reporting categories. They do not classify archived YouTube pages using CDX records,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table D.2: Literature groups reviewed for the 7+1 Archive-Side Evidence-State Model

Literature group	Main contribution	Limitation for this thesis
Platform governance and platform documentation	Defines platform-side states such as public, private, restricted, removed, disabled, demoted, or terminated [Gil18, Gil22, GBK20, Eur22, Goo26].	Does not classify archived YouTube pages using CDX records and replayable archived HTML.
Social-media persistence studies	Measures whether previously collected online content remains available, deleted, private, suspended, or otherwise unavailable [Zub18, MSM15].	Usually uses live recollection or dataset comparison, not archive-side replay evidence.
Web-archive coverage, redirection, and replay studies	Defines archive-related conditions such as no mementos, redirecting mementos, missing embedded resources, damaged replay, and retrieval bias [AAS+11, ANSVdS13, KAA+17, BKWN15, STvO+18, BKNW23].	Provides several relevant components, but not one integrated *7+1 archive-side evidence-state model*
Web-archive and digital-memory scholarship	Explains why archived records should be treated as partial, constructed, and dependent on access conditions [Bru18, BDA18].	Provides conceptual grounding, but not a technical archive-side evidence-state classifier.

replayable archived HTML, redirect-only records, archive gaps, or terminal absence. For this reason, they support the vocabulary of restriction and unavailability, but they do not provide the archive-side 7+1 archive-side evidence-state model used in this thesis.

D.5 Persistence and disappearance models

Persistence studies ask whether online items that were previously collected can still be retrieved later. These studies provide strong prior support for the general distinction between available and unavailable content.

Zubiaga (2018) [Zub18] studies the long-term persistence of Twitter datasets and measures how recollected datasets lose completeness and representativeness over time. This study is relevant because it shows that online items and their metadata may disappear or change after initial collection. Maddock et al. (2015) [MSM15] discuss missing tweets in historical Twitter data and distinguish deletion, private timelines, and deleted or suspended accounts as reasons why previously observed tweets may no longer be retrievable.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table D.3: Platform-side models and their relevance to the archive-side evidence states

Source type	States or categories	Closest thesis evidence states
YouTube documentation	Public, private, unlisted, age-restricted, region-restricted, locked-private, deleted, terminated.	Restricted; Unavailable.
Digital Services Act and DSA Transparency Database	Removal, disabling access, demotion, restriction of visibility, service suspension, account suspension, termination.	Restricted; Unavailable.
Platform governance literature	Content removal, access restriction, visibility reduction, labelling, account enforcement, automated moderation.	Restricted; Unavailable.

These studies support the general evidence states Available and Unavailable. They also provide partial support for Restricted, because private timelines or account-level restrictions limit access. They provide partial support for Eroded, because changing or missing metadata weakens the informational value of a previously collected record. However, these studies are not archive-side models. They do not classify CDX records, replayable archived HTML, redirect-only archive records, archive gaps, or terminal absence.

Table D.4 summarises this group.

Table D.4: Persistence and disappearance models relevant to the archive-side evidence states

Source	Classification or measured condition	Closest thesis evidence states
Zubiaga (2018) [Zub18]	Available and unavailable tweets; completeness; representativeness; similarity; changingness.	Available; Unavailable; partial support for Eroded.
Maddock et al. (2015) [MSM15]	Deleted by author, private timeline, deleted account, suspended account.	Restricted; Unavailable.

Table D.4 shows that persistence studies provide prior support for availability loss and access limitation, but they do not solve the archive-side classification problem of this thesis. They classify whether content can still be retrieved from a platform or dataset, not whether an archived YouTube page can be replayed, whether it redirects, whether it has lost contextual information, or whether the archive record contains a long gap.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

D.6 Archive-side coverage, redirection, and replay models

Web-archive literature is closest to the evidence problem of this thesis because it studies archived URLs, archived pages, redirects, missing resources, and replay limits.

Ainsworth et al. (2011) [AAS+11] examine how much of the web is archived and classify resources according to archive coverage, including resources with no archived copies and resources with one or more archived copies. This supports the need to distinguish archive presence from archive absence. However, archive coverage alone does not show whether an archived YouTube page displayed available, restricted, eroded, or unavailable content.

AlSum et al. (2013) [ANSVdS13] study archival HTTP redirection retrieval policies and distinguish between directly accessible archived records, redirecting archived records, mixed-status TimeMaps, and no-memento cases. Kelly et al. (2017) [KAA+17] also show that redirecting mementos can affect the count of usable archived representations. These studies provide strong archive-side support for separating redirect-only evidence from replayable archived page evidence.

Brunelle et al. (2015) [BKWN15] study memento damage and missing embedded resources. Their work shows that an archived page may exist but still be incomplete because some resources required for faithful replay are missing. Berlin et al. (2023) [BKNW23] study the transformation and replay of archived web pages and show that replay can preserve, reduce, or transform the original page experience. These studies support the thesis evidence state Eroded because they show that an archived page can exist while its informational value is weakened.

Samar et al. (2018) [STvO+18] study retrieval bias in web archive search and show that archived versions can be unevenly retrievable. Brügger (2018) [Bru18] treats the archived web as an object that must be studied as archived material rather than as a transparent copy of the live web. Ben-David and Amram (2018) [BDA18] also show that the Internet Archive participates in the socio-technical construction of historical records. These works support the thesis decision to preserve archive uncertainty instead of reducing all evidence to available or unavailable.

Table D.5 summarises the archive-side models and their relation to the archive-side evidence states.

Table D.5 shows that archive-side literature provides the strongest technical basis for the *7+1 archive-side evidence-state model*. It already discusses archive coverage, no archived copies, redirecting mementos, replay damage, missing resources, and retrieval limits. The limitation is that these components remain distributed across separate models and studies. They are not combined into one archive-side evidence-state model.

D.7 State-of-the-art limitation

The existing literature is sufficient to show that a simple binary model is not adequate for this thesis. However, it is not sufficient to provide a directly reusable *7+1 archive-side evidence-state model*.

A binary model would classify a video as either available or unavailable. This is too coarse for archive-side evidence. It would merge cases that must be kept separate. For example, a replayable archived page showing an unavailability message is not the same as a CDX record that only shows redirects. A long period without archive captures is not the same as a saved page that explicitly shows unavailability. A saved page with missing contextual information is not the same as a fully accessible saved page. These differences matter because the thesis measures what can be externally observed, not only whether the video is present or absent.

A CDX-only model is also insufficient. A CDX record can show that the archive recorded a URL at a specific time, but it does not by itself show what the saved YouTube page displayed. It cannot tell whether the page showed accessible content, a restriction message, missing metadata, or an unavailability notice. For this reason, CDX evidence must be separated from replayable archived HTML evidence.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table D.5: Archive-side models relevant to the archive-side evidence states

Source	Classification or measured condition	Closest thesis evidence states
Ainsworth et al. (2011) [AAS+11]	No archived copy, one archived copy, multiple archived copies, archive coverage.	Archive presence/absence; archived-copy coverage.
AlSum et al. (2013) [ANSVdS13]	HTTP 200 mementos, redirecting mementos, mixed-status TimeMaps, no mementos.	Available; Redirect-only; Terminal absence.
Kelly et al. (2017) [KAA+17]	Non-redirecting mementos, redirecting mementos, memento count differences.	Available; Redirect-only.
Brunelle et al. (2015) [BKWN15]	Memento damage, missing embedded resources, weighted replay damage.	Eroded.
Berlin et al. (2023) [BKNW23]	Replay transformation, replay style, reduced or transformed archived page experience.	Available; Eroded.
Samar et al. (2018) [STvO+18]	Retrieval bias and uneven retrievability of archived versions.	Retrieval bias; uneven discoverability; observability limits.

Platform-side models are also insufficient for the 7+1 archive-side evidence-state model. YouTube documentation and regulatory categories help define platform-side states such as private, restricted, removed, or terminated. These states support the conceptual distinction between restriction and unavailability. However, they do not classify archived HTML pages or archive index records. They describe platform-facing or regulatory states, not archive-side evidence states.

The state-of-the-art limitation is therefore specific. Existing models provide important components, but they do not provide the integrated archive-side classification model required by this thesis.

D.8 Novelty relative to prior literature

The *7+1 archive-side evidence-state model* is literature-grounded but thesis-defined. The individual phenomena are not all new. Available and unavailable content are established in persistence studies. Restriction is established in platform governance, platform documentation, and regulation. Metadata loss, replay damage, archived redirects, archive gaps, and no-memento cases are discussed in web-archive literature. What is thesis-specific is the integration of these phenomena into one 7+1 archive-side evidence-state model with one evidence logic.

Table D.6 compares each of the seven interpretable archive-side evidence states with its closest prior basis.

As Table D.6 shows, the seven interpretable archive-side evidence states have different degrees of prior support. Available, Restricted, and Unavailable have strong prior support as general availability or access states. Eroded and Redirect-only have strong archive-side technical antecedents, but their use as

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table D.6: Comparison of the seven interpretable archive-side evidence states with prior literature

Thesis evidence state	**Closest prior basis**	**Assessment**
Available	Available or retrievable items in persistence studies; HTTP 200 mementos in web-archive studies.	Strong prior basis. The thesis definition is narrower because it requires replayable archived HTML showing availability.
Restricted	Private, age-restricted, region-restricted, locked-private, disabled-access, and visibility-restricted states.	Strong platform-side basis. The thesis adapts restriction into an archive-side state when the restriction is visible in replayable archived HTML.
Eroded	Metadata change, memento damage, missing embedded resources, reduced replay quality.	Strong technical antecedents, but not usually defined as an archive-side evidence state. The thesis gives this condition an archive-side evidence label.
Unavailable	Deleted, removed, suspended, terminated, unavailable, or no longer retrievable content.	Strong prior basis. The thesis definition is narrower because it requires replayable archived HTML showing a platform-level unavailability notice.
Redirect-only	Redirecting mementos and archived HTTP redirection.	Close archive-side antecedent. The thesis uses this condition as a separate archive-side evidence state when no replayable archived HTML is available.
Archival gap	Archive coverage; no archived copy; one archived copy; multiple archived copies.	Partial prior basis. Prior literature supports archive coverage and archived-copy presence or absence, but it does not define Archival gap as a distinct archive-side evidence state. The evidence state is defined in this thesis as a time-gap condition within the 7+1 archive-side evidence-state model.
Terminal absence	No archived copy; archive coverage; no mementos.	Partial prior basis. Prior literature supports archive absence and no-memento conditions, but it does not define Terminal absence as a distinct archive-side evidence state. The evidence state is defined in this thesis as the absence of further replayable archived HTML after the last usable saved page.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

archive-side evidence states is thesis-specific. Archival gap and Terminal absence are related to archive coverage and loss of retrievability, but their specific roles in the *7+1 archive-side evidence-state model* are defined in this thesis.

The novelty claim is therefore bounded:

> No prior model was identified in the reviewed literature that defines the same seven interpretable archive-side evidence states plus one residual unknown archive evidence state with the same evidence rules. The contribution of the thesis is the integration of prior concepts into one reproducible 7+1 archive-side evidence-state model for YouTube availability evidence.

This statement does not claim proof of non-existence. It states the result of the reviewed literature.

D.9 Evidence model

The thesis uses two evidence layers. The first is the Internet Archive layer. The second is the YouTube platform layer. These layers are related, but they have different roles.

The Internet Archive layer provides archive-side evidence. This includes CDX records, capture timestamps, HTTP status codes, file types, redirect records, replayable archived HTML pages, time gaps, and the absence of later replayable pages. The archive-side evidence states are derived from this layer.

The YouTube platform layer provides current platform-facing evidence. This includes current YouTube responses, endpoint-level status, platform-facing error categories, and HTML-derived playability signals at the time of measurement. This layer is not used to assign the archive-side evidence states. It is used separately to describe current platform-facing availability, to compare current platform evidence with archive evidence, and to support later statistical analysis.

Table D.7 summarises the two layers.

Table D.7: Evidence layers and their role in the thesis

Evidence layer	**Signals covered**	**Used for evidence states?**	**Role in the thesis**
Internet Archive layer	CDX records, capture timestamps, HTTP status codes, file types, redirect records, replayable archived HTML pages, time gaps, and absence of later replayable pages.	Yes	Defines the archive-side evidence states and supports over-time archive-based measurement.
YouTube platform layer	Current platform-facing response, endpoint status, platform error category, and current HTML-derived playability signal.	No	Describes current YouTube availability, supports comparison with archive evidence, and contributes to later statistical analysis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table D.7 is important because it prevents a common misunderstanding. The overall thesis uses both archive-side and platform-facing evidence, but the *7+1 archive-side evidence-state model* itself is derived from archive-side evidence. Platform-facing signals are not used to assign the archive-side evidence states.

D.10 Stepwise derivation process

This section explains how raw archive-side evidence is converted into the archive-side evidence states. The process begins with one YouTube video identifier and then reads the available archive evidence for that identifier.

First, the system checks whether any CDX record exists for the video URL. If no CDX record exists, the archive provides no archive-side trajectory for that identifier, and no archive-side evidence state is assigned. If CDX evidence exists, the system then checks whether replayable archived HTML exists.

The second step separates index evidence from page evidence. Index evidence means that a CDX record exists. Page evidence (replayable page) means that the archived HTML page can be opened and inspected. This distinction is necessary because only a replayable archived HTML page can show what the saved YouTube page displayed.

The third step applies page-based classification. If a replayable archived HTML page exists, the system inspects what the page shows. If the page shows accessible video content, the state is Available. If the page shows an access limitation, the state is Restricted. If the page exists but has lost important contextual information, the state is Eroded. If the page shows a platform-level unavailability message, the state is Unavailable.

The fourth step applies index-based classification. If no replayable archived HTML page exists, but the CDX records show redirect behaviour, the state is Redirect-only. This state is separate from Unavailable because no saved page is available to show an unavailability message.

The fifth step checks for long periods without archive captures. If there is a long period without captures, the state is Archival gap. This does not mean that the video was unavailable during that period. It means only that the archive provides no evidence during that period.

The sixth step checks the end of the archive record. If no later replayable archived HTML page appears after the last usable saved page, the state is Terminal absence. This does not prove that the video disappeared at that exact time. It means that the usable archived page record ends.

Table D.8 summarises the evidence-to-state rules.

Table D.8 shows that the evidence-state names are not arbitrary labels. Each evidence state is tied to a specific evidence condition.

D.11 Evidence-state naming logic

The evidence-state names describe observable evidence conditions. They do not describe hidden YouTube causes.

The name Available is used when the archived page shows access to the video page. The name Restricted is used when the archived page shows an access limitation. The name Eroded is used when the archived page still exists, but important contextual information is missing or degraded. The name Unavailable is used when the archived page shows a platform-level unavailability message. The name Redirect-only is used when the archive records redirects but no inspectable saved page is available. The name Archival gap is used when no archive captures exist for a defined period. The name Terminal absence is used when no later replayable archived HTML page appears after the last usable saved page.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table D.8: Evidence-to-state derivation

Evidence condition	**Evidence state**	**Reason for separate state**
Replayable archived HTML page shows the video page as accessible.	Available	The saved page provides direct evidence of archived accessibility.
Replayable archived HTML page shows an access limitation.	Restricted	The saved page exists, but access is limited.
Replayable archived HTML page exists, but important contextual information is missing or degraded.	Eroded	The saved page is not absent, but its informational value is reduced.
Replayable archived HTML page shows a platform-level unavailability message.	Unavailable	The saved page provides explicit evidence of unavailability.
CDX records show redirects, but no replayable archived HTML page is available.	Redirect-only	The archive records technical interaction, but no page-level meaning can be inspected.
The archive record has a long period without captures.	Archival gap	The archive provides no evidence during that period.
No later replayable archived HTML page appears after the last usable saved page.	Terminal absence	The usable saved-page record ends.
CDX or processed snapshot evidence exists, but the available evidence is insufficient for assignment to one of the seven interpretable archive-side evidence states.	Unknown archive evidence	The archive evidence exists, but it does not support a substantive evidence-state interpretation.

This naming rule keeps the 7+1 archive-side evidence-state model tied to external evidence. It avoids unsupported claims about platform intent, moderation causes, or internal removal reasons.

D.12 Split-and-merge rationale for the evidence-state model

The structure of the 7+1 archive-side evidence-state model follows from a split-and-merge rationale. A separate archive-side evidence state is retained when three conditions are met:

1. the evidence condition is observable from the thesis data;
2. the evidence condition has a different meaning for the research question;

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3. merging it with another archive-side evidence state would remove information that the thesis needs to preserve.

A separate evidence state is not retained when a difference is only a minor technical variant and does not change the meaning of the evidence. For example, the model does not create a separate evidence state for every HTTP status code, every wording of an error message, every missing metadata field, or every possible gap length.

Table D.9 documents this rationale.

Table D.9: Split-and-merge decisions in the archive-side evidence-state model

Candidate decision	**Decision**	**Reason**
Merge explicit unavailability, redirect-only records, archive gaps, and terminal absence into one missing category.	Rejected	These conditions have different evidence meanings. A saved page saying unavailable is not the same as a redirect-only record or a missing capture period.
Create separate evidence states for every HTTP status code.	Rejected	HTTP status codes are technical signals, but not all status-code differences create a separate page-level meaning for this thesis.
Create separate evidence states for each wording of a platform error message.	Rejected	Different wording may express the same evidence condition. Splitting by wording would reduce reproducibility.
Create separate evidence states for each missing metadata field.	Rejected	Missing title, description, or channel information are retained as evidence details under the broader evidence state Eroded.
Separate replayable archived HTML from CDX-only evidence.	Retained	Replayable archived HTML can be inspected. CDX-only evidence cannot show page-level meaning.
Separate archive gaps from explicit unavailability.	Retained	A gap means that the archive provides no evidence. It does not show that the video was unavailable.

Table D.9 shows why the model contains seven interpretable archive-side evidence states plus one residual unknown archive evidence state. Fewer evidence states would merge evidence conditions that have different meanings. More evidence states would mainly split technical variants that do not support a separate research claim.

D.13 Boundary of completeness

The *7+1 archive-side evidence-state model* is complete only within the defined archive-side evidence input space. This means that completeness is not a claim about all possible YouTube behaviours. It is a claim

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

about the evidence that the classifier is allowed to read.

The boundary has three parts. First, the archive-side evidence states are derived only from external Wayback evidence. This includes CDX records, capture timestamps, redirect records, replayable archived HTML pages, capture gaps, and the absence of later replayable archived HTML. Live YouTube platform signals and YouTube-internal information are not used to assign these archive-side evidence states.

Second, each archive-side evidence state must represent a distinct observable evidence condition. A separate archive-side evidence state is retained only when merging it with another archive-side evidence state would hide a relevant difference in the evidence. Minor technical variants are not treated as new evidence states unless they change the meaning of the evidence for RQ1.

Third, the 7+1 archive-side evidence-state model must be checked against the archive evidence used in the thesis. Every observed archive case must either fit one of the seven interpretable archive-side evidence states or be retained as a residual unknown archive evidence state. If the data repeatedly show a new archive evidence case that is visible, relevant to RQ1, and not covered by any existing evidence state, then the 7+1 archive-side evidence-state model is not complete and must be revised.

The 7+1 archive-side evidence-state model does not identify internal YouTube causes such as uploader deletion, copyright enforcement, policy enforcement, account suspension, automated moderation, manual moderation, regional blocking, or technical failure unless such a cause is directly visible in the external evidence. The model therefore avoids causal claims that the evidence cannot support.

The completeness claim is procedural. It concerns the measurement model and its input evidence. It does not concern the full internal reality of the platform.

D.14 Empirical population note

The seven interpretable archive-side evidence states plus one residual unknown archive evidence state define the possible archive-side evidence states of the classifier. This does not mean that all seven interpretable archive-side evidence states must appear equally in the empirical data. Some archive-side evidence states may be rare or absent in a specific corpus because the archive evidence may not contain the signals required to assign them.

It is also necessary to distinguish the formal evidence-state space from boundary categories. Unknown archive evidence is the residual unknown archive evidence state of the 7+1 archive-side evidence-state model and is part of the formal model as U_{unk}.

Table D.10 summarises this distinction.

Table D.10: Formal evidence-state space and empirical observation categories

Category type	**Meaning**	**Example**
Formal evidence state	A state defined by the *7+1 archive-side evidence-state model* and tied to an evidence rule.	Available; Redirect-only; Archival gap; Unknown archive evidence.
Boundary category	A reporting category outside the 7+1 archive-side evidence-state model.	No archival trajectory.
Observed distribution	The actual frequency of evidence states or boundary categories in the corpus.	Many available cases; few restricted cases.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table D.10 prevents a second common misunderstanding. The formal 7+1 archive-side evidence-state model defines what the classifier can distinguish. The empirical distribution shows what was actually observed in the corpus.

D.15 Relation to RQ1 and RQ3

The *7+1 archive-side evidence-state model* supports RQ1 by defining what counts as a measurable archive-side evidence state. RQ1 does not identify all possible platform-side deletion or suppression causes. It reports the evidence states that can be measured from the archive-side evidence used in this thesis.

The 7+1 archive-side evidence-state model also supports RQ3 because it shows why a preservation-oriented measurement system should not reduce all evidence to available or unavailable. A binary label would lose information about restriction, metadata erosion, redirect-only evidence, archive gaps, and terminal absence. The *7+1 archive-side evidence-state model* therefore supports a system design that records evidence limits instead of hiding them.

The relation to RQ1 and RQ3 is summarised in Table D.11.

Table D.11: Relation of the 7+1 Archive-Side Evidence-State Model to the research questions

Research question	**Role of the 7+1 Archive-Side Evidence-State Model**
RQ1	Defines the measurable archive-side evidence states used to report bounded patterns of access evidence, restriction evidence, degradation evidence, unavailability evidence, redirects, gaps, terminal absence, and residual unknown archive evidence state.
RQ3	Supports the design and evaluation of a preservation-oriented measurement artefact by showing why evidence limits and partial preservation states must be recorded explicitly.

D.16 Summary

The state of the art provides several partial models relevant to online content availability. Platform-side models define restriction, removal, and account-level enforcement. Persistence studies define available and unavailable content and show that metadata and dataset completeness can change over time. Web-archive studies define archive coverage, redirecting mementos, replay damage, missing resources, and retrieval bias. Digital memory and digital forgetting literature explains why weakened access and partial preservation matter.

These prior models provide the basis for the seven interpretable archive-side evidence states, but they do not provide the same integrated 7+1 archive-side evidence-state model. The *7+1 archive-side evidence-state model* is therefore literature-grounded but thesis-defined. Its purpose is to classify externally observable archive evidence in a reproducible way. Its scope is limited to the evidence model used in this thesis. It does not claim to identify all real-world YouTube removal causes, and it does not claim that no further platform-side patterns exist.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

APPENDIX E

Derivation and Literature Positioning of the RQ3 Completeness Artefact

E.1 Aim and scope

This appendix explains how the system-level implications for RQ3 were derived and how the completeness artefact is positioned relative to prior work on digital preservation, web archives, and archive observability.

RQ3 asks which system-level design improvements can support digital memory preservation under external observability limits. In this appendix, system-level improvements do not mean general policy recommendations for YouTube or claims about how platforms should moderate content. They mean design requirements for an external observational measurement system. The system should record incomplete evidence explicitly, distinguish different levels of archival preservation, avoid unsupported causal claims, and produce reproducible per-video outputs.

The appendix has five aims. First, it defines the technical terms used in the RQ3 completeness artefact. Second, it reviews prior models and methods relevant to preservation completeness and observability limits. Third, it explains why existing models are not sufficient for the specific measurement problem of this thesis. Fourth, it documents the derivation chain from observed limits to system-level design implications. Fifth, it defines the stopping rule that determines when the RQ3 derivation is complete within the evidence model used in this thesis.

The appendix follows literature-review guidance that a review should map existing work, classify approaches, identify limitations, and make clear how a new contribution is grounded in prior literature [Sny19, BCK14]. The purpose is not to prove that no similar model exists anywhere. The purpose is to show that, within the reviewed state of the art, no prior model was identified that provides the same public-evidence, item-level representation used in this thesis: *Absent*, *Trace*, *Partial*, and *Full* preservation levels, combined with a bounded scalar completeness score and explicit missingness.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

E.2 Plain definitions of technical terms

Table E.1 defines the technical terms used in this appendix.

Table E.1: Basic terms used in the RQ3 appendix

Term	Plain definition
Preservation completeness	The degree to which a video identifier is represented by archival evidence. In this thesis, completeness is based on whether CDX records, replayable HTML, and WARC-level payload evidence are observable.
Observability limit	A boundary in what the researcher can know from external evidence. For example, a CDX record may show that the archive indexed a URL, but it does not show what the archived YouTube page displayed.
System-level improvement	A design implication for the measurement system. It describes how the system should record and expose evidence limits. It is not a recommendation about YouTube moderation policy.
CDX record	An archive index record showing that a URL was captured or indexed at a specific time. It may expose fields such as timestamp, MIME type, HTTP status code, filename, offset, or record length, depending on the archive interface.
Replayable HTML	An archived HTML page that can be opened and inspected through the replay interface. Replayable HTML provides stronger evidence than a CDX record because the saved page content can be read.
WARC payload	The underlying archived payload stored in a WARC file. In the completeness model, WARC-level payload evidence represents the strongest preservation layer because it can provide access to the archived record behind the replayed representation.
Semantic signal	Recoverable YouTube-specific meaning inside archived or replayed evidence, such as a player response or playability status. A semantic signal is not the same as the mere existence of an archived page.
Missingness	An explicitly recorded absence of evidence. Missingness is not repaired by guessing and is not treated as proof that the video disappeared or was removed.

Two distinctions are central. First, index evidence and replay evidence are different. A CDX record shows that the archive has an index-level trace, but it does not by itself show page content. Second, archival observability and platform behaviour are different. A completeness score describes how much archival evidence is externally observable. It does not explain why a video is available, unavailable, restricted, or missing on YouTube.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

E.3 State of the art for preservation completeness and observability limits

Before deriving system-level improvements, it is necessary to check which prior models and methods already represent preservation completeness or observability limits. The relevant state of the art is distributed across several model families.

OAIS defines digital preservation as a system-level archival responsibility and distinguishes submission, archival, and dissemination information packages [Con24b]. PREMIS provides a preservation-metadata vocabulary for objects, events, agents, rights, and preservation-related metadata [PRE15]. Repository trustworthiness and maturity models, including NDSA Levels, TRAC/ISO 16363, and CoreTrustSeal, evaluate preservation programmes, repository governance, and organisational capability [Nat19, Con24a, Cor23]. Web-archive formats, protocols, and access layers provide different parts of the technical evidence base. WARC stores archived web records, Memento supports time-based access to prior resource states, CDX exposes archive index traces, and WACZ packages WARC data, indexes, and metadata for distribution and replay [Int20, Int25a, Web23, VdSNS13]. Item-level web-archive research examines replay damage, correspondence, archivability, dynamic-web replay problems, retrieval bias, and archival voids [BKWN15, RA18, WNAG23, STvO+18, AWN21]. Practitioner tools such as Archive-It QA and Browsertrix QA operationalise parts of this problem through missing-resource checks, patch crawls, and replay-focused quality assurance [Arc24, Web25].

Table E.2 summarises how these model families relate to the RQ3 completeness artefact.

The reviewed state of the art shows that preservation is not a binary problem. Existing work already provides repository-level preservation models, preservation metadata frameworks, web-archive formats, access protocols, replay-quality measures, and retrieval-bias methods. However, these approaches do not provide the exact model required by this thesis: a public-evidence, item-level completeness artefact for YouTube-related Wayback evidence that combines index presence, replayable HTML, WARC-level evidence, semantic-signal availability, snapshot count, temporal span, and explicit missingness.

E.4 State-of-the-art limitation

The limitation of prior work is not that preservation models are absent. The limitation is more specific. Existing models do not provide one directly reusable artefact for representing how complete the archival evidence is for each individual video identifier in the corpus.

Repository-level models such as OAIS, NDSA Levels, TRAC/ISO 16363, and CoreTrustSeal are too broad for the thesis measurement problem. They assess archival systems, preservation programmes, repository responsibilities, or trustworthiness criteria. They are useful for understanding digital preservation as a system-level activity, but they do not tell the researcher whether a specific YouTube video has no archival trace, only CDX evidence, replayable HTML, or WARC-level payload evidence.

Preservation-metadata models such as PREMIS are also not directly sufficient. PREMIS can record preservation events, objects, rights, and provenance. However, this type of metadata is usually available inside repository systems, not through public Wayback interfaces. The thesis cannot assume access to internal preservation-event metadata.

Web-archive formats and protocols are closer to the thesis problem. WARC, CDX, WACZ, and Memento provide ways to package, index, discover, and access archived web material. Yet they remain technical substrates. They expose evidence but do not define how to convert that evidence into a bounded completeness level and score.

Replay-damage, correspondence, dynamic-web replay, retrieval-bias, and archival-void studies are closest to the item-level problem. They show that replayed pages can be incomplete, temporally inconsistent, difficult to retrieve, or absent from an archive [BKWN15, RA18, WNAG23, STvO+18, AWN21]. These

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table E.2: Prior model families relevant to preservation completeness and observability limits

Model family	Main unit	Contribution	Limitation for this thesis
OAIS	Archive or preservation system	Defines long-term preservation as an archival system with information packages and preservation responsibilities [Con24b].	Does not provide a public, item-level completeness state for a single archived YouTube page.
PREMIS	Digital object, event, agent, rights metadata	Provides a vocabulary for recording preservation metadata, provenance, preservation events, and rights information [PRE15].	Assumes preservation metadata that is usually internal to repositories and not available through public Wayback evidence.
Repository trust and maturity models	Repository or preservation programme	Assess organisational trustworthiness, storage, integrity, governance, and preservation capability [Nat19, Con24a, Cor23].	The unit of analysis is the repository or programme, not one video identifier or one archived page.
Web-archive formats and access protocols	Capture record, archive package, URI version	WARC, CDX, WACZ, and Memento expose capture records, payload structure, timestamps, and time-based discovery [Int20, Int25a, Web23, VdSNS13].	They expose evidence surfaces, but they do not themselves define an interpretive completeness taxonomy.
Replay quality and damage models	Replayed page or memento	Replay-damage research measures the effect of missing embedded resources on replay quality [BKWN15].	It captures one important failure mode, but not the broader CDX–HTML–WARC completeness structure used here.
Correspondence and archivability models	Archived page or website	Quality research distinguishes correspondence, relevance, and archivability as dimensions of web-archive quality [RA18].	These concepts support non-binary assessment, but they do not provide the thesis-specific per-video scoring rule.
Dynamic-web replay and retrieval-bias studies	Replay process, search result, URI space	These studies show that replay, search, and discovery can introduce hidden observability limits [WNAG23, STvO+18, AWN21].	They describe important limits, but they do not combine them into a public-evidence item-level completeness artefact.
Practitioner QA tools	Crawl result or replay session	Archive-It QA and Browsertrix QA detect missing resources, replay problems, and crawl/replay differences [Arc24, Web25].	They are operational tools rather than a portable theoretical completeness model for analysis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

works justify treating observability limits as first-class analytical conditions. However, each addresses a specific part of the problem. None of them provides the combined per-video model used in this thesis.

The state-of-the-art limitation is therefore as follows:

> Prior work provides important components for representing preservation and observability limits, but no reviewed model provides the same public-evidence, item-level completeness artefact used in this thesis: ordinal states for Absent, Trace, Partial, and Full preservation, a bounded scalar score, explicit missingness, and compatibility with per-video statistical analysis.

This is a bounded novelty claim. It does not claim that no similar model exists anywhere. It states that no directly reusable model was identified within the reviewed state of the art.

E.5 Position of the thesis artefact

The completeness artefact is literature-grounded but thesis-defined. It uses prior work to justify three design decisions.

First, it treats preservation as layered rather than binary. This is consistent with digital preservation and web-archive literature, which shows that preservation depends on more than the existence of a URL or file [Con24b, PRE15, Int20, VdSNS13].

Second, it treats public archive evidence as limited. Web-archive evidence may show capture presence, index records, or replayable pages, but this does not guarantee semantic recoverability or full reconstruction [BKWN15, WNAG23, STvO+18].

Third, it treats absence and missingness as analytical information. Missing archive traces, missing replayable HTML, missing WARC-level access, and missing semantic signals are not silently ignored. They are recorded as part of the measurement output.

The thesis-specific contribution is the integration of these ideas into a reproducible per-video artefact. The artefact represents each video identifier through a completeness level, a scalar completeness score, and explicit evidence indicators.

E.6 Derivation chain for RQ3

The RQ3 system-level implications are derived through a traceable chain. They are not generated as open-ended recommendations. Table E.3 summarises this chain.

The derivation chain is important because it answers the feedback question: the system-level improvements are not invented freely. They come from observed limits in the measurement system, are represented in the completeness artefact, and are evaluated against predefined criteria.

E.7 Observed preservation and observability limits

The completeness artefact is derived from observed limits in the archival evidence. These limits are listed in Table E.4.

These observed limits define the empirical basis of the RQ3 artefact. They show that preservation in this corpus is not simply present or absent. It is layered, partial, and unevenly observable.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table E.3: Derivation chain for RQ3 system-level design implications

Step	Input	Output for RQ3
Research gaps	The thesis identifies limits in external observability, heterogeneous evidence, non-binary trace states, and reproducibility.	Defines the problem that the system-level artefact must address.
Observed limits	The empirical system observes absent archive traces, CDX-only traces, replayable HTML without full payload evidence, missing semantic signals, uneven snapshot counts, and uneven temporal spans.	Defines which preservation and observability limits must be represented.
Completeness artefact	The artefact encodes preservation levels, a bounded scalar score, and explicit evidence indicators.	Turns observed limits into fixed variables.
Evaluation criteria	Criteria C1–C5 test observability differentiation, preservation of partial evidence, non-causal use, reproducibility, and alignment with the problem framing.	Determines whether the artefact satisfies the RQ3 requirements.
System-level implications	Only implications grounded in the observed limits and evaluated artefact properties are retained.	Produces the final RQ3 design implications.

E.8 Completeness artefact as response to the observed limits

The completeness artefact converts observed preservation limits into explicit variables. For each video identifier v, the artefact produces:

- `completeness_level`,
- `completeness_score`,
- `has_cdx`,
- `has_html`,
- `has_warc_payload`,
- `has_semantic_signal`,
- `snapshot_count`,
- `span_days`.

The four completeness levels are shown in Table E.5.

The completeness level records the strongest observable preservation layer. The scalar score then refines the level by adding bounded information about archive density and temporal coverage.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table E.4: Observed preservation and observability limits used for RQ3

Observed limit	**Plain meaning**	**Representation in the artefact**
No archival trace	Some videos have no CDX record, no replayable HTML, and no recoverable semantic signal.	Represented as *Absent* with modal score 0.0.
CDX-only trace	The archive index records the URL, but no saved page can be inspected.	Represented as *Trace* with modal score 0.3.
Replayable HTML without full preservation	A saved HTML page can be opened, but full WARC-level payload evidence is not available.	Represented as *Partial* with modal score 0.7.
Full WARC-level preservation not observed	The formal model defines *Full* as CDX evidence, replayable HTML, and WARC-level payload evidence. In the analysed corpus, no case reaches this level because the required payload-level evidence is not observed.	Represented through the *Full* level, modal score 1.0, and `has_warc_payload`; empirically observed with zero cases.
Missing semantic signals	A record may exist without recoverable YouTube-specific semantic evidence.	Represented through `has_semantic_signal`.
Uneven snapshot count	Some videos have more archived observations than others.	Represented through a bounded snapshot-count component.
Uneven temporal span	Some videos have archive evidence spread over longer periods than others.	Represented through a bounded temporal-span component.
Missing or unclear inputs	Some evidence cannot be interpreted safely.	Missingness is preserved rather than imputed.
Completeness is not a deletion signal	Low completeness does not mean that a video was removed or suppressed.	The artefact is interpreted as archival observability, not platform behaviour.
Completeness is not an exclusion rule	Videos with low completeness remain part of the corpus where analytically appropriate.	The artefact conditions interpretation rather than filtering the dataset.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table E.5: Completeness levels used in the RQ3 artefact

Level	Required evidence	Modal score	Plain meaning
Full	CDX + replayable HTML + WARC payload	1.0	The strongest defined preservation condition. Index evidence, replayable representation, and payload-level evidence are all available.
Partial	CDX + replayable HTML	0.7	The page can be replayed and inspected, but full payload-level preservation is not available.
Trace	CDX only	0.3	The archive recorded an index-level trace, but no replayable archived page is available.
Absent	No archival record	0.0	No archive-side trace is observable for the video identifier.

The score is calculated as:

$$S_v = 0.60 \cdot M_v + 0.25 \cdot C_v + 0.15 \cdot T_v,$$

where:

- M_v is the modal score of the completeness level,
- $C_v = \min\left(\frac{\text{snapshot_count}_v}{10}, 1\right)$ is the normalised snapshot-count component,
- $T_v = \min\left(\frac{\text{span_days}_v}{180}, 1\right)$ is the normalised temporal-span component.

The score is bounded by the modal preservation state:

$$S_v = \min\left(0.60 \cdot M_v + 0.25 \cdot C_v + 0.15 \cdot T_v, M_v\right).$$

This bounding rule prevents snapshot count or temporal span from compensating for a missing preservation layer. A video with only CDX evidence cannot become *Partial* only because it has many CDX records. A video without WARC-level evidence cannot become *Full* only because it has a long archive span.

E.9 Coverage check against the thesis evidence

The completeness artefact was checked against the evidence conditions observed in the thesis data. The purpose of this check is not to repeat the empirical results reported in Chapter 5. Its purpose is to show that the artefact can represent the preservation limits that occurred in the evidence model.

The observed evidence conditions include cases with no archival trace, CDX-only trace evidence, replayable HTML without full payload-level evidence, missing semantic signals, uneven snapshot counts, uneven temporal spans, and missing or unclear inputs. These conditions are represented by the artefact through the preservation levels, the bounded scalar score, explicit evidence indicators, and missingness rules.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The empirical results reported in Tables 5.14, 5.15, and 5.16 show that the artefact was applied to the analytical corpus and that the observed preservation conditions are not binary. In particular, the results distinguish absent archive evidence, CDX-only trace evidence, replayable HTML evidence, and missing semantic signals. The absence of empirically observed *Full* preservation does not remove *Full* from the model; it shows that this level is part of the formal evidence model but was not reached in the analysed corpus.

This coverage check is part of the stopping rule for RQ3. The artefact would require revision if the thesis data repeatedly showed a preservation or observability limit that changed the interpretation of archival evidence and could not be represented by a preservation level, the bounded score, an evidence indicator, or a missingness rule.

E.10 Constraints of the RQ3 derivation

The RQ3 derivation is bounded by five constraints.

External evidence constraint. The artefact uses only externally observable archival and replay-layer evidence. It does not use YouTube-internal data, moderation logs, platform intent, or private repository metadata.

Preservation-layer constraint. Each completeness level must represent a distinct preservation condition. *Absent*, *Trace*, *Partial*, and *Full* are separated because they describe different evidence layers.

Score-bounding constraint. Snapshot count and temporal span may refine the completeness score within a preservation level, but they cannot compensate for a missing preservation layer.

No-guessing constraint. Missing WARC evidence, missing replayable HTML, missing semantic signals, and absent archive traces are not filled by assumption. Missing evidence remains visible.

No-causal-claim constraint. Completeness does not explain why a video is unavailable, restricted, deleted, suppressed, or missing. It describes archival observability only.

These constraints prevent the artefact from making claims that exceed the evidence.

E.11 Stopping rule for RQ3

The stopping rule for RQ3 answers the question of when the derivation of system-level improvements is complete.

For RQ3, completeness does not mean that no further system-level improvements are possible in the world. That cannot be proven. Completeness means that all observed preservation and observability limits within the thesis evidence model are handled explicitly by the artefact.

A preservation or observability limit is handled when it is represented in one of five ways:

1. by a completeness level;
2. by the bounded scalar score;
3. by an explicit missingness or evidence indicator;
4. by a documented non-use constraint;

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5. by one of the evaluation criteria C1–C5.

The model is not considered complete if the data repeatedly show a preservation limit that affects archival observability and is not represented by any of these mechanisms. In that case, the artefact would need revision.

A plain formulation is:

> The RQ3 derivation is complete when every repeated preservation limit observed in the evidence is either represented by the completeness artefact or explicitly excluded from interpretation because the evidence does not support a design claim. The thesis does not claim that all possible system-level improvements are exhausted. It claims only that the proposed improvements cover the observed limits of the evidence model used in this study.

E.12 Evaluation criteria and derived system-level implications

The system-level implications are evaluated through criteria C1–C5. Table E.6 maps each criterion to the corresponding design implication.

Table E.6: Evaluation criteria and derived RQ3 system-level implications

Criterion	What is evaluated	Derived system-level implication
C1	The artefact must distinguish different states of archival observability.	Use layered observability states instead of a binary available/unavailable label.
C2	The artefact must preserve partial traces, indeterminate states, and incomplete evidence.	Retain CDX-only evidence, partial replay evidence, missing semantic signals, and absent traces instead of discarding them.
C3	The artefact must remain analytically usable without causal attribution.	Keep archival completeness separate from platform intent, moderation cause, deletion reason, or content value.
C4	The artefact must be reproducible and transparent.	Compute predefined per-video variables through fixed rules that can be rerun and checked.
C5	The artefact must remain aligned with the thesis problem framing.	Expose infrastructural limits of digital memory without embedding normative rankings or unsupported content valuations.

The resulting system-level implications are therefore not independent recommendations. They are derived from observed evidence limits and evaluated artefact properties.

E.13 Boundary of the RQ3 claim

The RQ3 completeness artefact has a bounded claim.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

It does not claim that all preservation problems are solved. It does not claim that all possible system-level improvements have been identified. It does not claim that low completeness means deletion, suppression, or low importance. It does not claim that high completeness means uninterrupted YouTube availability or high historical value.

The artefact claims only that, for the evidence used in this thesis, archival observability can be represented through four preservation levels, a bounded scalar score, and explicit evidence indicators. This allows the thesis to report incomplete preservation as an empirical condition rather than treating it as noise, error, or grounds for exclusion.

E.14 Summary

The reviewed state of the art provides several important components for representing preservation and observability limits. OAIS and PREMIS support the general archival and metadata framing. Repository maturity and trustworthiness models support systematic preservation assessment. WARC, CDX, WACZ, and Memento provide the technical substrate for archive evidence. Replay-damage, correspondence, dynamic-web replay, retrieval-bias, and archival-void research show that public archive evidence is partial and uneven. Practitioner QA tools show how some of these limits can be operationalised in workflows.

However, these components are distributed across different model families. They do not provide the same public-evidence, item-level completeness artefact required by this thesis. The thesis therefore integrates prior concepts into a bounded artefact that distinguishes Absent, Trace, Partial, and Full preservation, computes a bounded scalar score, and preserves missingness explicitly.

This appendix answers the RQ3 feedback point as follows: the system-level improvements are derived from observed preservation and observability limits in the measurement system, not from speculation. The derivation is complete only when those observed limits are represented by the artefact or explicitly bounded by non-use constraints. The contribution is therefore methodological and evidence-bound.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX F

Taxonomy Mapping Tables

This appendix documents the axis-specific mapping used in the human taxonomy feature construct. Table F.1 presents the English raw values, their mapped semantic groups, the corresponding semantic group identifiers, and the retained-row status. To avoid redundant glossary references, each semantic group identifier is shown only at its first occurrence within the table.

Table F.1: Full taxonomy hierarchy from raw values to semantic groups and retained rows

Axis	Field	Raw value	Semantic group	Semantic group ID	Ret.
A	MC	Armed violence / combat aftermath	Armed violence and combat	axis_A_main_SEM01	TRUE
A	MC	Art / graffiti	Artistic and symbolic expression	axis_A_main_SEM09	FALSE
A	MC	Bombardment / shelling	Armed violence and combat		TRUE
A	MC	Civilian displacement / refugee movement	Displacement and humanitarian crises	axis_A_main_SEM08	TRUE
A	MC	Clashes / combat	Armed violence and combat		TRUE
A	MC	Demonstration / protest	Protest and civic mobilisation	axis_A_main_SEM02	TRUE
A	MC	Destruction / damage	Armed violence and combat		TRUE
A	MC	Funeral / mourning	Human casualties and losses	axis_A_main_SEM03	TRUE
A	MC	Humanitarian / relief activity	Humanitarian activity	axis_A_main_SEM06	TRUE
A	MC	Interview / talk / media format	Media and journalistic content	axis_A_main_SEM07	FALSE

Continued on next page

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Axis	Field	Raw value (EN)	Semantic group (EN)	Semantic group ID	Ret.
A	MC	Mass killing / massacre	Human casualties and losses		TRUE
A	MC	Medical treatment / rescue	Humanitarian activity		TRUE
A	MC	Political statement / conference	Political activity	axis_A_main_SEM05	TRUE
A	MC	Raid / arrest / detention	Repression and violations	axis_A_main_SEM04	TRUE
A	MC	Refugee camp / humanitarian hardship	Displacement and humanitarian crises		TRUE
A	MC	Security-force repression	Repression and violations		TRUE
B	GRP	Civilian	Civilian	axis_B_group_SEM04	TRUE
B	GRP	Humanitarian	Humanitarian	axis_B_group_SEM06	TRUE
B	GRP	Martyrs	Martyrs and fatalities	axis_B_group_SEM03	TRUE
B	GRP	Military	Military	axis_B_group_SEM02	TRUE
B	GRP	Reports and statements	Reports and statements	axis_B_group_SEM05	FALSE
B	GRP	Violations	Violations	axis_B_group_SEM01	TRUE
C	SUB	Air strike	Armed violence and combat	axis_C_subcat_SEM02	TRUE
C	SUB	Arrest / detention	Repression and violations	axis_C_subcat_SEM06	TRUE
C	SUB	Armed presence / patrol	Organisational or military activity	axis_C_subcat_SEM04	TRUE
C	SUB	Art / song / poem	Artistic and cultural expression	axis_C_subcat_SEM09	FALSE
C	SUB	Assault / beating	Repression and violations		TRUE
C	SUB	Burial / funeral	Human casualties and losses	axis_C_subcat_SEM01	TRUE
C	SUB	Checkpoint / control	Organisational or military activity		TRUE
C	SUB	Children / women / civilians (descriptive)	Non-event subject-referential content	axis_C_subcat_SEM11	FALSE
C	SUB	Demonstration	Protest and civic mobilisation	axis_C_subcat_SEM05	TRUE
C	SUB	Documentary / report / interview	Media and journalistic activity	axis_C_subcat_SEM07	FALSE
C	SUB	Funeral procession	Human casualties and losses		TRUE
C	SUB	Geographic place reference	Spatial context	axis_C_subcat_SEM08	FALSE
C	SUB	Injury / wounded	Human casualties and losses		TRUE
C	SUB	Killing / martyrdom	Human casualties and losses		TRUE
C	SUB	Military training / coordination	Organisational or military activity		TRUE
C	SUB	Mortar / shelling	Armed violence and combat		TRUE

Continued on next page

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Axis	Field	Raw value (EN)	Semantic group (EN)	Semantic group ID	Ret.
C	SUB	Political initiative / institution	Political activity	axis_C_subcat_SEM10	FALSE
C	SUB	Property destruction / looting	Repression and violations		TRUE
C	SUB	Protest strike / sit-in	Protest and civic mobilisation		TRUE
C	SUB	Raid / incursion	Armed violence and combat		TRUE
C	SUB	Sniper fire / shooting	Armed violence and combat		TRUE
C	SUB	Suppression of protest	Repression and violations		TRUE
C	SUB	Temporal marker / anniversary	Temporal context	axis_C_subcat_SEM03	FALSE
C	SUB	Weapons and military materiel	Weapons and military materiel	—	FALSE
C	SUB	Political expression	Political expression	—	FALSE

Abbreviations: MC = `main_classification`; GRP = `group`; SUB = `subcategory`; Ret. = Row retained. Semantic group IDs are shown only at first occurrence. Two Axis C groups have no uploaded glossary entry and are therefore shown as -- rather than inferred.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Figures

3.1 DSR-architecture of the observational measurement system. 36

4.1 Distribution of CDX snapshot counts per video (log-scaled frequency). 76
4.2 Distribution of non-zero archival lifespans in days (log-scaled x-axis). 77
4.3 Distribution of multi-day gap counts per video for videos with at least one gap (> 1 day). 78
4.4 Year of last observed CDX snapshot per video. Grey bars indicate right-censored final observation years. 79
4.5 Distribution of archived HTML snapshots per video. The distribution is strongly right-skewed: most videos are represented by a single snapshot, while a small subset exhibits substantially higher snapshot counts. 79
4.6 Distribution of implemented snapshot-level classifier labels across all snapshot observations. 85
4.7 Distribution of observed snapshot-level classifier-label transition counts per video. 86
4.8 Proportion of snapshot observations preceded by a CDX temporal gap, by snapshot-level classifier label. 87
4.9 Observed YouTube availability descriptor outcomes. The figure contrasts videos with an explicitly observed terminal unavailable state and descriptors that remain right-censored at the end of the observation period. 87
4.10 Log-scaled cross-source alignment matrix between live YouTube platform states and archived HTML semantic states. Cells indicate the number of videos for each state combination. Logarithmic scaling reveals sparse but systematic misalignments that are visually suppressed in linear representations. 90
4.11 Videos without prior archival HTML semantic evidence before live YouTube checks. 91

5.1 Distribution of observed archival unavailability and cases with no observed archival unavailability in the RQ1 analytical population. 102
5.2 Binned distribution of CDX temporal gap counts among videos with Wayback history. Gap counts use the implemented one-day threshold and are grouped to show the zero-heavy distribution and the long right tail without excluding cases. 104

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Tables

1.1 Alignment between research gaps (G1–G4) and thesis contributions 6

2.1 Overview of how the present thesis engages with the claims of the Vienna Manifesto on Digital Humanism [WLA+19]. Each cell names where the claim frames, constrains, informs, or is discussed within the thesis; the section mark (§) links to where the alignment is documented. Claim identifiers link to their glossary definitions. FC = Foundational Claim; P = Principle. 19

3.1 Primary Data Sources and Their Analytical Purpose 26
3.2 Operationalisation of completeness from observable archival evidence to preservation level and ordinal score. 60
3.3 Alignment of methodological elements with research questions (RQ1–RQ3). 71

4.1 Core descriptive statistics for Wayback CDX temporal measures. 75
4.2 Selected proportions describing Wayback CDX coverage and discontinuities. . . . 75
4.3 Coverage of Wayback HTML snapshots relative to the CDX snapshot universe . . 78
4.4 Structural characteristics of archived HTML snapshots 80
4.5 Composition of archived HTML snapshots by player response 80
4.6 Presence of YouTube playability signals in archived HTML snapshots 81
4.7 Presence of video-related metadata fields in archived HTML snapshots 81
4.8 Distribution of YouTube platform availability states observed at the time of measurement 82
4.9 Transport-level characteristics of platform observations classified as `unknown` . . . 83
4.10 Video-level summary of archived availability descriptors derived from Wayback evidence. 84
4.11 Intersection between Wayback CDX records and archived HTML snapshots. . . . 88
4.12 Semantic yield of archived HTML snapshots with exact CDX timestamp matches. 89
4.13 Joint distribution of live YouTube platform states and archived HTML semantic states. 90
4.14 Videos with at least one prior archival HTML semantic signal before live YouTube checks. 91
4.15 Human taxonomy label coverage across the analytical corpus 92
4.16 Reduction of field-specific taxonomy values into retained grouped categories . . . 92
4.17 Structural consistency between human labels and semantic-top features 93
4.18 Retained grouped categories by taxonomy axis 94
4.19 Cross-axis semantic-top presence per video 94

5.1 RQ1 dataset summary: corpus size, platform availability, and archival observability. 98
5.2 RQ2 content-feature availability before outcome filtering 99

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.3 Analytical datasets and their role in Chapter 5 . . . 100
5.4 RQ1 analytical population and archival observability subsets . . . 101
5.5 Distribution of YouTube platform availability states within the RQ1 analytical population . . . 101
5.6 Video-level cross-tabulation of HTML-derived YouTube unavailability and absence of Wayback history, grouped by endpoint availability label . . . 101
5.7 Observed versus unobserved unavailability within the RQ1 analytical population . . . 102
5.8 Archival presence and absence within the RQ1 analytical population . . . 102
5.9 Non-exclusive video-level evidence membership across archive-side evidence states 103
5.10 Archival observability boundary conditions within the RQ1 analytical population . 103
5.11 CDX temporal fragmentation among videos with Wayback history . . . 104
5.12 Video-level construction of the RQ2 regression population from HTML-derived YouTube availability outcomes . . . 105
5.13 Selected semantic content attributes and their conditional associations with final platform-observed availability outcomes ($\lambda = 0.3$). . . . 106
5.14 Distribution of ordinal archival completeness levels . . . 108
5.15 Summary statistics for completeness score . . . 108
5.16 Cross-tabulation of archival completeness level and semantic-signal availability . . 109
5.17 Objects evaluated in the Design Cycle evaluation reported in Chapter 5 . . . 110
5.18 Functional-testing evidence for the observational measurement system . . . 111
5.19 Evaluation of system support for RQ1 . . . 112
5.20 Evaluation of system support for RQ2 . . . 113
5.21 Evaluation of system support for RQ3 . . . 113
5.22 Structured criteria mapping of the observational measurement system against C1–C5 114
5.23 Discrepancy log and remaining evaluation boundaries . . . 115

6.1 E1–E3 mapped to framework emphases (short clauses). . . . 120

A.1 Overview of generative AI tools used . . . 131

B.1 Construct dictionary derived from literature grounding . . . 133
B.2 Variable list derived from the observational pipeline . . . 134
B.3 Methodological precedents supporting the literature grounding . . . 135

D.1 Basic terms used in the appendix . . . 143
D.2 Literature groups reviewed for the 7+1 Archive-Side Evidence-State Model . . . 144
D.3 Platform-side models and their relevance to the archive-side evidence states . . . 145
D.4 Persistence and disappearance models relevant to the archive-side evidence states 145
D.5 Archive-side models relevant to the archive-side evidence states . . . 147
D.6 Comparison of the seven interpretable archive-side evidence states with prior literature 148
D.7 Evidence layers and their role in the thesis . . . 149
D.8 Evidence-to-state derivation . . . 151
D.9 Split-and-merge decisions in the archive-side evidence-state model . . . 152
D.10 Formal evidence-state space and empirical observation categories . . . 153
D.11 Relation of the 7+1 Archive-Side Evidence-State Model to the research questions 154

E.1 Basic terms used in the RQ3 appendix . . . 156
E.2 Prior model families relevant to preservation completeness and observability limits 158
E.3 Derivation chain for RQ3 system-level design implications . . . 160
E.4 Observed preservation and observability limits used for RQ3 . . . 161

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

E.5 Completeness levels used in the RQ3 artefact 162
E.6 Evaluation criteria and derived RQ3 system-level implications 164

F.1 Full taxonomy hierarchy from raw values to semantic groups and retained rows . . 167

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Algorithms

1 Observed Archive-Side Evidence-State Classifier for a Single Video 48

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Glossary

$t_{\mathbf{avail}}$ The first archival timestamp at which a video is observed as available based on HTML-derived playability evidence.. 101

$t_{\mathbf{unavail}}$ The first archival timestamp at which a video is observed as explicitly unavailable based on HTML-derived playability evidence.. 101

Account terminated Account terminated refers to a live, endpoint-derived observation associated with a video whose associated account is reported as terminated in the platform-facing outcome.. 82

Age-restricted Age-restricted refers to a live, endpoint-derived observation indicating that access to a video is restricted due to an age-related constraint.. 82

Analysis stage Analysis stage refers to the workflow stage at which no additional filtering, transformation, or manual correction is applied to the produced data artefacts.. 73

Application Programming Interface A standardised protocol enabling software systems to exchange data, such as the YouTube Data API or the Internet Archive CDX API used in this study.. 189

Archival gap An archive-side evidence state interval exceeding a defined long-gap threshold during which no archival snapshots are recorded despite prior or subsequent archival evidence. This evidence state is distinct from the one-day CDX temporal gap count used as a structural fragmentation metric.. 103

Archival observability The condition under which a video's availability or unavailability can be externally observed through archival records such as CDX entries or replayable HTML evidence.. 100

Archival trajectory The existence of at least one archival interaction record (e.g. CDX snapshot, HTML replay evidence) for a given video identifier within a web archive. A video with an archival trajectory has entered the archive's observational scope, regardless of whether the archived content is complete, partial, or later disappears.. 102

Archived HTML snapshots Archived HTML snapshots refers to archived replayed HTML pages associated with a video identifier at a specific archival timestamp and analysed at snapshot level by the system.. 37, 45, 51

Archived payloads Archived payloads refers to archived content payloads referenced by snapshot metadata records and stored by the system as payload artefacts when retrieved, including file-path and integrity metadata.. 74

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Availability descriptors (longitudinal) Availability descriptors (longitudinal) refers to reconstructed per-video descriptor records that summarise longitudinal observation histories as archive-side evidence state intervals with onsets and durations derived deterministically from snapshot evidence and system mapping rules.. 83

Availability trajectory A per-video, temporally ordered representation of externally observable availability-related signals derived from platform-facing responses and archival evidence. Availability trajectories encode observed availability, unavailability, and right-censoring without inferring platform intent or causal mechanisms.. 34, 44, 94, 98

Available Available refers to a live, endpoint-derived observation indicating platform-facing availability at the time of measurement.. 21, 82

Available An archive-side evidence state in which an archived copy of the YouTube page shows that the video page was accessible at that archived time.. 103

axis_A_main_SEM01 Armed violence and combat. Content depicting or documenting armed confrontation, military operations, shelling, airstrikes, or other forms of organised physical violence, including their immediate material effects such as destruction or damage.. 167

axis_A_main_SEM02 Protest and civic mobilisation. Content relating to demonstrations, protests, and other forms of collective civic mobilisation, including both peaceful and confrontational public gatherings organised to express political or social demands.. 167

axis_A_main_SEM03 Human casualties and losses. Content focusing on human casualties resulting from violence, including deaths, injuries, mass killings, and the recovery or documentation of victims following violent events.. 167

axis_A_main_SEM04 Repression and violations. Content documenting acts of coercion, repression, or rights violations, including arrests, security operations, dispersal of protests, and other practices associated with the enforcement of control through force or intimidation.. 168

axis_A_main_SEM05 Political activity. Content relating to formal or informal political activity, including statements, organisational announcements, press conferences, and political coordination not directly depicting violent action.. 168

axis_A_main_SEM06 Humanitarian activity. Content related to humanitarian response and relief activities, such as medical assistance, emergency aid, rescue operations, and organised efforts to support affected civilian populations.. 167

axis_A_main_SEM07 Media and journalistic content. Content primarily characterised by journalistic or media formats, such as interviews, reports, opinion polling, or discussion programmes, in which the communicative form is foregrounded over the direct documentation of events.. 167

axis_A_main_SEM08 Displacement and humanitarian crises. Content documenting forced displacement, refugee movements, and broader humanitarian crises, including conditions of mass displacement, loss of shelter, and large-scale civilian hardship.. 167

axis_A_main_SEM09 Artistic and symbolic expression. Content expressing events or experiences through artistic, cultural, or symbolic forms, including visual art, performances, or other non-documentary representations.. 167

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

axis_B_group_SEM01 Violations. Content focused on alleged or documented violations, including abuses, transgressions, or misconduct directed against individuals or groups.. 168

axis_B_group_SEM02 Military. Content in which military actors, armed forces, or military operations constitute the primary reference frame.. 168

axis_B_group_SEM03 Martyrs and fatalities. Content centred on individuals identified as martyrs or fatalities, including commemoration, identification, or documentation of death.. 168

axis_B_group_SEM04 Civilian. Content in which civilians or civilian contexts are the primary reference, including civilian life, civilian spaces, and non-combatant populations.. 168

axis_B_group_SEM05 Reports and statements. Content consisting primarily of formal reports, statements, or informational bulletins where declarative communication is foregrounded rather than event documentation.. 168

axis_B_group_SEM06 Humanitarian. Content framed around humanitarian conditions, needs, or assistance, including relief activities and support for affected populations.. 168

axis_C_subcat_SEM01 Human casualties and losses. Content focusing on human losses resulting from violence, including deaths, injuries, funerals, mourning practices, and documentation of victims.. 168

axis_C_subcat_SEM02 Armed violence and combat. Content depicting direct armed confrontation or combat activities, including attacks, shelling, air strikes, sniper fire, and other forms of organised physical violence.. 168

axis_C_subcat_SEM03 Temporal context. Content defined primarily by temporal framing rather than by an event or action.. 169

axis_C_subcat_SEM04 Organisational or military activity. Content related to military or organisational operations that are preparatory, logistical, or structural in nature such as patrols, training, coordination, or territorial control.. 168

axis_C_subcat_SEM05 Protest and civic mobilisation. Content documenting protests, strikes, demonstrations, sit-ins, or other forms of collective civic or political mobilisation.. 168

axis_C_subcat_SEM06 Repression and violations. Content documenting acts of repression, coercion, or violations against civilians or opposition actors including assaults, property destruction, and suppression of dissent.. 168

axis_C_subcat_SEM07 Media and journalistic activity. Content primarily characterised by media formats such as interviews, reports, talk shows, documentaries, or informational programming.. 168

axis_C_subcat_SEM08 Spatial context. Content defined primarily by geographic or locational reference rather than by an event or action.. 168

axis_C_subcat_SEM09 Artistic and cultural expression. Content expressing experiences, ideas, or narratives through artistic, cultural, or symbolic forms such as music, poetry, visual art, performance, or creative media.. 168

axis_C_subcat_SEM10 Political activity. Content related to organised political processes or governance-oriented actions such as initiatives, institutions, or political projects.. 169

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

axis_C_subcat_SEM11 Non-event subject-referential content. Content that is descriptive or contextual in nature and does not document a specific event, action, or occurrence.. 168

Capture Index file format A structured index maintained by the Internet Archive that records metadata about archived web captures, including URL, timestamp, MIME type, and checksum identifiers.. 189

Causal inference A class of analytical approaches that aim to attribute causal relationships or mechanisms; treated in this thesis as out of scope for externally observed availability patterns.. 33

Collective memory A socially constructed record of shared experiences and narratives that shape a community's understanding of its past.. 1, 7, 21

Cross-layer observability patterns Recurring configurations of externally observable signals across multiple evidence layers used in this thesis, including CDX index records, archived HTML representations, and archived payload records. These patterns are used to support classification and comparison under platform opacity, not to infer platform-internal causes.. 17

Cross-source alignment / non-alignment / discrepancies Cross-source alignment / non-alignment / discrepancies refers to agreement or divergence between signals derived from CDX records, archived HTML-derived evidence, and live platform check outcomes for the same video identifiers, as reported in this chapter.. 89

Data artefacts Data artefacts refers to the data outputs produced by the deterministic observational measurement system and used as the sole empirical basis for analysis in this chapter.. 73

Design Science Research (DSR) A research framework that combines analytical inquiry with the construction and evaluation of research artefacts in response to a defined problem context, structured here through relevance, build, and evaluation cycles.. 21, 189

Digital amnesia A collective-level reduction in the accessible digital record of past events that occurs when digital content becomes unavailable, partially accessible, or inconsistently archived across platforms and web archives. In this thesis, digital amnesia results from cumulative instances of digital forgetting that limit the practical retrievability of digital records over time.. 9, 26, 123

Digital forgetting A set of gradual processes through which previously accessible digital records lose usable access over time. In this thesis, this includes loss of visibility, loss of accessibility, or loss of contextual information, and it can result in partial persistence such as index-only traces or fragmented archival captures rather than a single, complete deletion event.. 7, 8, 21

Digital Humanism An interdisciplinary framework advocating the alignment of digital technologies with human values, transparency, accountability, and societal well-being.. 22

Empirical boundary for platform observability Empirical boundary for platform observability refers to the boundary condition describing what can be empirically observed under the combined constraints of CDX records, archived HTML replay evidence, and live platform check outcomes used in this chapter.. 81

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Eroded An archive-side evidence state in which an archived copy of the YouTube page exists, but important context such as title, description, or channel information is missing or degraded.. 103

Error Error refers to a live, endpoint-derived observation indicating failure to obtain a determinate platform-facing state, excluding explicitly rate-limited cases classified as Unknown.. 82

Explicitly unavailable An HTML-derived YouTube availability outcome in which the YouTube page data contains a clear signal showing that the video was unavailable at the time of the check. This excludes rate limits, missing signals, and other unclear cases.. 101, 105

Indeterminate An HTML-derived YouTube availability outcome in which the system could not determine whether the video was available or unavailable because the YouTube page data did not contain a clear availability signal.. 105

Joint observability Joint observability refers to what can be co-observed across CDX records, archived HTML replay evidence, and live platform check outcomes for the same video identifiers and observation points, including conditions under which co-observation fails.. 88

Live platform checks / YouTube platform checks Live platform checks / YouTube platform checks refers to point-in-time, per-video observations obtained via live requests and stored as one record per `(run_id, video_id)` in the live-check results table.. 37

Logarithmic frequency scale / logarithmic scaling Logarithmic frequency scale / logarithmic scaling refers to the use of logarithmic transformation in plotting to visualise highly skewed frequency distributions across multiple orders of magnitude.. 75

Manifesto Foundational Claim 1 The disruption of societal institutions and political structures by the flood of data and algorithms is the defining challenge of the digital age [WLA+19].. 19, 25

Manifesto Foundational Claim 3 Digital technologies are shaped by implicit and explicit choices that incorporate values, norms, and economic interests; many of these choices remain hidden in software programs implementing algorithms that remain invisible [WLA+19].. 19

Manifesto Principle 1 Both sciences and humanities need to be involved in designing, developing and deploying digital systems. This calls for genuine interdisciplinary engagement from the earliest stages of technological development [WLA+19].. 18

Manifesto Principle 10 New educational curricula are needed at all levels of education to prepare citizens and professionals for the challenges of the digital age [WLA+19].. 18

Manifesto Principle 11 Early computer science and informatics education needs to be reformed to include ethical, social, and humanistic perspectives alongside technical skills [WLA+19].. 18

Manifesto Principle 2 Platforms need to be altered to better safeguard the free expression of opinion, the dissemination of information, and the protection of privacy [WLA+19].. 15, 16, 19, 123

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Manifesto Principle 3 Effective regulations must ensure prediction accuracy, fairness, accountability, and transparency of software programs and algorithms [WLA+19].. 17, 19, 63, 69

Manifesto Principle 4 Market concentration in the digital economy needs to be reduced. Antitrust authorities need to be strengthened and empowered to act against monopolistic practices [WLA+19].. 18

Manifesto Principle 5 Decisions with an impact on people's lives should be made by accountable humans, not algorithms. Digital systems should enhance, not substitute, human judgment [WLA+19].. 17–19, 118, 119, 123

Manifesto Principle 6 Scientific approaches crossing different disciplines are a prerequisite for tackling the challenges ahead. Technological disciplines must collaborate with social sciences, humanities, and other sciences [WLA+19].. 16, 19, 25

Manifesto Principle 7 Universities must create an environment that fosters interdisciplinary research and cooperation between computer science, the social sciences, and the humanities [WLA+19].. 18

Manifesto Principle 8 Academic and industrial researchers must engage openly with wider society and reflect upon their approaches [WLA+19].. 19, 25, 69

Manifesto Principle 9 No technology is neutral. Designers, developers, and deployers must acknowledge their shared responsibility for the impact of information technologies [WLA+19].. 15, 19, 21

Meta-preservation A process in which digital infrastructures not only preserve content but also record and maintain information about the act of preservation itself - such as capture events, timestamps, and omissions.. 64

Methodological instrument An implemented observational system designed to enable structured, reproducible measurement and analysis of externally observable states, without intervening in platform governance, policy, or content operations.. 73

No archival trajectory A condition in which a video identifier has no corresponding archival records in the web archive and therefore cannot support longitudinal availability analysis.. 102

Observability The extent to which the state, availability, or properties of a digital artefact can be empirically determined through externally accessible signals such as platform-facing responses, archival indices, or replayable representations. In this thesis, observability is strictly limited to what can be measured without access to platform-internal or archive-internal states.. 82, 97, 122

Observational pattern A measurable regularity identified across the RQ1 analytical population based exclusively on externally observable platform and archival evidence.. 104

Observational window The finite time interval during which archival observations are available for a given video identifier, bounded by the first and last recorded archival interactions. Events occurring outside this window are unobservable.. 77, 101

Platform governance regimes The combination of platform rules, policy frameworks, and enforcement practices that determine what content is allowed, how it is made visible, and how compliance is enforced, including moderation systems. In this thesis, governance regimes are treated as a source of instability for long-term content availability.. 7

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Platform opacity The condition under which platform-internal states, decision processes, or moderation actions are not accessible to external observers. In this thesis, platform opacity constrains empirical claims to externally observable signals only.. 8, 21, 24

Platform unavailability A platform-facing condition in which a video is not accessible on YouTube at the time of observation, independent of archival coverage.. 100

Playability signals Playability signals refers to the system's archived-HTML evidence indicators consisting of whether a structured player response was extracted and, when present, the extracted playability status and associated reason fields.. 50–53

Playable An HTML-derived YouTube availability outcome in which the YouTube page data contains a clear player signal showing that the video was available at the time of the check. This is a system classification and does not mean that the video was manually watched.. 105

Player response Player response refers to the structured YouTube player response object extracted by the system from archived HTML using defined extraction carriers, and it serves as the source object for playability signal fields when extraction succeeds.. 50–53, 80, 173

Processual Describing a phenomenon that develops over time through multiple steps rather than occurring as a single event. In this thesis, the term is used to characterise digital forgetting and changes in availability as time-extended and multi-stage.. 7, 21

Redirect-only An archive-side evidence state in which the archive repeatedly records redirects instead of an archived YouTube video page that can be inspected.. 103

Replayable HTML snapshot An archived web page for which the archival replay system returns an HTML document that can be retrieved and parsed at a given snapshot timestamp. Replayability in this thesis refers only to the delivery of an HTML representation and does not imply that embedded media, scripts, or video content can be executed or played.. 41, 88, 118, 125

Response shape Response shape refers to the categorical classification assigned by the system to an archived HTML response to describe the observed response type, including distinctions between captured site HTML, replay-layer wrappers or notices, and transport or decoding outcomes.. 53

Restricted An archive-side evidence state in which an archived copy of the YouTube page shows an access limit, such as age restriction, regional restriction, or required login.. 103

Right censoring A condition in time-to-event analysis where no archival unavailability event is observed within the available archival observation window.. 98, 101

Right-censored descriptor An availability trajectory for which no terminal unavailable state is observed within the finite observation window. In this thesis, right-censoring indicates that disappearance cannot be empirically confirmed, and is treated as an explicit empirical outcome rather than as missing or incomplete data.. 84

RQ1 analytical population The set of videos included in RQ1, defined as all video identifiers for which an availability descriptor could be reconstructed from archival evidence, with platform-facing checks analysed separately. 101

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Sampling frame The explicitly defined set of units and selection rules that determine which identifiers enter the observational system and thereby condition all subsequent analyses.. 27

Semantic observability The extent to which an archived or replayed representation exposes interpretable, platform-specific meaning—such as availability status, playability signals, or contextual metadata—required to classify the state of a digital artefact. Semantic observability is distinct from structural completeness and may be absent even when a page is technically valid.. 81

Semantic presence within archived HTML Semantic presence (within archived HTML) refers to the presence of recoverable YouTube-specific semantic evidence in an archived HTML snapshot, operationalised in this chapter via successful extraction of playability signals from a player response.. 81

Semantic yield (archived HTML) Semantic yield (archived HTML) refers to the extent to which archived HTML snapshots expose recoverable YouTube-specific semantic evidence under the system's extraction logic, operationalised via the extraction outcomes and stored semantic-signal indicators.. 52, 53, 88

Snapshot An archived capture of a web resource at a specific point in time, represented by a CDX index record that references the corresponding archived record in an ARC/WARC file (via filename, byte offset, and length).. 45, 74

Snapshot timestamps Snapshot timestamps refers to timestamp identifiers used to index snapshot-level archive observations and to support per-video temporal reconstruction.. 74

Societal-scale relevance Relevance that concerns collective practices, shared meanings, and socially consequential processes beyond individual use. In this thesis, the term is used to justify treating platform-hosted Syria-related content as part of the environment in which collective memory can be formed, maintained, or disrupted.. 15, 16

Socio-technical Describing outcomes shaped jointly by technical components, such as software, storage, ranking systems, and archive infrastructure, and social or institutional components, such as policies, moderation practices, and user practices. In this thesis, both platform behaviour and observed remembering and forgetting dynamics are treated as socio-technical.. 8, 37, 65

Socio-technical systems Systems in which technical infrastructure and social or institutional arrangements are integrated and jointly shape observed outcomes. In this thesis, digital platforms are explicitly treated as socio-technical systems that reshape social processes and outcomes.. 15

Structural fragmentation The presence of CDX temporal discontinuities within a video's archival record, indicating non-continuous observability without implying content disappearance during the discontinuity interval.. 104

Structural observability boundary Structural observability boundary refers to the boundary described in this chapter where archived HTML snapshots lack structural or semantic evidence prior to live platform checks, with structural evidence operationalised by the system's HTML integrity indicators and semantic evidence operationalised by extraction outcomes.. 52, 53

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Technological neutrality The view that digital technologies are passive instruments that do not reshape social reality. In this thesis, this view is rejected, and platforms are treated as socio-technical systems that reshape social processes and outcomes.. 14

Temporal gap Temporal gap refers to a CDX-level interval greater than one day between consecutive archival snapshots of a video. It is used as a structural fragmentation metric and is distinct from the Archival gap evidence state.. 76

Terminal absence An archive-side evidence state in which an earlier archived YouTube page showed a clear state, later archive records still exist, but no later archived YouTube page shows a clear state again.. 103

Unavailable An archive-side evidence state in which an archived copy of the YouTube page shows a YouTube message that the video is unavailable.. 103

Unknown Unknown refers to a residual live category used when platform-facing availability cannot be determined; in this dataset it corresponds to unsuccessful requests associated with HTTP 429 and the internal HTTP error code `rate_limited`, and it is treated as an unobservable platform state rather than evidence of availability, restriction, or removal.. 82

Unknown archive evidence A residual unknown archive evidence state in which CDX or processed snapshot evidence exists, but the available evidence is insufficient to assign the observation to one of the seven interpretable archive-side evidence states.. 103, 124, 126

Video identifier Video identifier refers to the primary unit key used by the system to associate observations with a specific YouTube video, stored as `video_id` in system tables.. 74

Video-related metadata fields Video-related metadata fields / video-level context refers to snapshot-level contextual metadata indicators recorded by the system for archived HTML snapshots, including boolean indicators for the presence of channel-related fields and title or description fields.. 81, 173

Wayback CDX snapshot metadata / CDX records Wayback CDX snapshot metadata / CDX records refers to snapshot-level metadata records used as the baseline description of the archival snapshot universe for each video identifier, including timestamp values, HTTP status codes, MIME type strings, and references to archived payloads.. 37, 50, 74

Wayback Machine An Internet Archive service providing access to historical web snapshots via CDX indexes and replayable HTML or WARC content.. 64, 74

Wayback Machine replay layer Wayback Machine replay layer / Wayback replay refers to the replay mechanism through which archived resources are retrieved as rendered HTML responses, including replay-layer wrapper HTML and notices distinct from captured site content.. 37

Web ARChive file format A Web ARChive container format storing HTTP request–response data and payloads for archived web resources.. 189

Wrapper-based nature of Wayback replay Wrapper-based nature of Wayback replay refers to the property that replay-layer wrapper HTML can be structurally valid regardless of whether the underlying capture yields semantically informative content, and the system distinguishes such wrapper outputs via response-shape classification.. 80

YouTube A video-sharing platform whose availability states form the empirical basis of this analysis.. 10, 11, 15, 16, 24, 123

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acronyms

API application programming interface. 74

CDX capture index file format. 26, 32, 33, 35, 38, 39, 45, 60, 63, 64, 66, 68, 100, 102, 108, 118, 119

CRISP-DM Cross-Industry Standard Process for Data Mining. 32, 63

DSR Design Science Research (DSR). 19, 21–23, 27, 28, 32, 33, 35, 66, 70, 71

HTML Hypertext Markup Language. 11

IA Internet Archive. 39

MIME Multipurpose Internet Mail Extensions. 38, 39, 45, 64, 74

NGO Non-Governmental Organization. 26, 73

NLP natural language processing. 55, 56, 58, 93

PRIMAD Provenance, Reference, Instrument, Method, Actor, Data. 19, 32, 33, 63, 67–69

URL Uniform Resource Locator. 39

WARC web archive file format. 33, 38, 60, 63–65, 108, 119

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

[AAS+11] Scott G. Ainsworth, Ahmed Alsum, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson. How much of the web is archived? In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 133–136, New York, NY, USA, June 2011. Association for Computing Machinery.

[Aas19] Susan Aasman. Finding traces in YouTube's living archive: Exploring informal archival practices. *TMG – Journal for Media History*, 22(1):35–55, 2019.

[Adr24] Robbert-Jan Adriaansen. Latent and explicit mnemonic communities on social media: studying digital memory formation through hashtag co-occurrence analysis. *Memory, Mind & Media*, 3:e13, January 2024.

[AJAKD+19] Abdul Rahman Al Jaloud, Hadi Al Khatib, Jeff Deutch, Dia Kayyali, and Jillian C. York. Caught in the net: The impact of "extremist" speech regulations on human rights content. Technical report, Electronic Frontier Foundation and Syrian Archive and WITNESS, May 2019.

[AK20] Amelia Acker and Adam Kreisberg. Social media data archives in an API-driven world. *Archival Science*, 20(2):105–123, June 2020.

[Alk26] Amer Alkojjeh. Digital Amnesia Observational Measurement System — Thesis Release, February 2026. Version v1.1-thesis. Archived release DOI: 10.5281/zenodo.19429629.

[ANSVdS13] Ahmed AlSum, Michael L. Nelson, Robert Sanderson, and Herbert Van de Sompel. Archival HTTP redirection retrieval policies. In *Proceedings of the temporal web analytics workshop*, 2013.

[Arc24] Archive-It. How to patch crawl with the wayback QA tool, 2024.

[Ass11] Jan Assmann. *Cultural memory and early civilization*. Cambridge University Press, Cambridge, 2011.

[AWN21] Sawood Alam, Michele C. Weigle, and Michael L. Nelson. Profiling web archival voids for memento routing. In *Proceedings of the 2021 ACM/IEEE joint conference on digital libraries (JCDL)*, pages 150–159, 2021. arXiv: 2108.03311 [cs.DL].

[Bar07] Karen Barad. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press, 2007.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[BCK14] Sebastian K. Boell and Dubravka Cecez-Kecmanovic. A Hermeneutic Approach for Conducting Literature Reviews and Literature Searches. *Communications of the Association for Information Systems*, 34, 2014.

[BDA18] Anat Ben-David and Adam Amram. The Internet Archive and the socio-technical construction of historical facts. *Internet Histories*, 2(1-2):179–201, April 2018.

[BDMN24] Anat Ben-David, Oren Meyers, and Motti Neiger. How social memory works on social media: A methodological framework. *Memory, Mind & Media*, 3:e20, January 2024.

[BKNW23] John Berlin, Mat Kelly, Michael L. Nelson, and Michele C. Weigle. To re-experience the web: a framework for the transformation and replay of archived web pages. *ACM Transactions on the Web*, 17(3), 2023.

[BKWN15] Justin F. Brunelle, Mat Kelly, Michele C. Weigle, and Michael L. Nelson. Not all mementos are created equal: Measuring the impact of missing resources. *International Journal on Digital Libraries*, 16(3–4):283–301, 2015.

[BM23] Lucas M. Bietti and Eric Mayor. A longitudinal study of conversational remembering in WhatsApp group messages before, during, and after COVID-19 lockdown. *Memory, Mind & Media*, 2:e5, January 2023.

[Bra13] Rosi Braidotti. *The posthuman*. Polity, Cambridge, 2013.

[Bru18] Niels Bruegger. *The archived web: Doing history in the digital age*. MIT Press, Cambridge, MA, 2018.

[CD21] Keith Cortis and Brian Davis. Over a decade of social opinion mining: a systematic review. *Artificial Intelligence Review*, 54(7):4873–4965, October 2021.

[Com25] Common Crawl Foundation. Common crawl: Open web data for research, 2025.

[Con24a] Consultative Committee for Space Data Systems. Audit and certification of trustworthy digital repositories. Recommended Practice CCSDS 652.0-M-2, Consultative Committee for Space Data Systems, December 2024.

[Con24b] Consultative Committee for Space Data Systems. Reference model for an open archival information system (OAIS). Recommended Practice CCSDS 650.0-M-3, Consultative Committee for Space Data Systems, December 2024.

[Cor23] CoreTrustSeal Standards and Certification Board. CoreTrustSeal trustworthy data repositories requirements 2023–2025, 2023.

[DH18] Jeff Deutch and Hadi Habal. The Syrian Archive: A Methodological Case Study of Open-Source Investigation of State Crime Using Video Evidence from Social Media Platforms. *State Crime Journal*, 7:46, April 2018.

[Eur] European Commission. DSA transparency database.

[Eur22] European Parliament and Council of the European Union. Regulation (EU) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/EC (digital services act), October 2022.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[Eur26] European Commission. Research API – DSA transparency database, 2026.

[FFJ+16] Nicola Ferro, Norbert Fuhr, Kalervo Järvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. Increasing reproducibility in IR: Findings from the dagstuhl seminar on "reproducibility of data-oriented experiments in e-Science". *SIGIR Forum*, 50(1):68–82, June 2016.

[Flo14] Luciano Floridi. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press, Oxford, New York, June 2014.

[GBK20] Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big data & society*, 7(1):205395171989794–, 2020.

[GH13] Shirley Gregor and Alan R. Hevner. Positioning and presenting design science research for maximum impact1. *MIS Quarterly*, 37(2):337–355, 2013.

[Gil18] T. Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[Gil22] Tarleton Gillespie. Do not recommend? Reduction as a form of content moderation. *Social Media + Society*, 8(3), 2022.

[Goo26] Google Developers. Videos | YouTube data API | google for developers, 2026.

[Hay17] N. Katherine Hayles. *Unthought: The Power of the Cognitive Nonconscious*. University of Chicago Press, 2017.

[HC10] Alan Hevner and Samir Chatterjee. *Design Research in Information Systems: Theory and Practice*, volume 22 of *Integrated Series in Information Systems*. Springer US, Boston, MA, 2010.

[Hev07] Alan R. Hevner. A three cycle view of design science research. *Scandinavian Journal of Information Systems*, 19(2):87–92, 2007.

[HGA16] Helge Holzmann, Vinay Goel, and Avishek Anand. ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 83–92, Newark New Jersey USA, June 2016. ACM.

[HKS+15] Hugo C. Huurdeman, Jaap Kamps, Thaer Samar, Arjen P. de Vries, Anat Ben-David, and Richard A. Rogers. Lost but not forgotten: finding pages on the unarchived web. *International Journal on Digital Libraries*, 16(3):247–265, September 2015.

[HMPR04] Alan Hevner, Salvatore March, Jinsoo Park, and Sudha Ram. Design Science in Information Systems Research. *Management Information Systems Quarterly*, 28(1), March 2004.

[Hos17] Andrew Hoskins. *Digital memory studies*. Routledge, 2017.

[Hub22] Hillary Hubley. Bad Speech, Good Evidence: Content Moderation in the Context of Open-Source Investigations. *International Criminal Law Review*, 22(5-6):989–1015, January 2022.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[Int20] Internet Archive. Information and documentation. WARC file format, 2020.

[Int25a] Internet Archive. CDX server API for the wayback machine, 2025.

[Int25b] Internet Archive. Wayback machine: Web archiving platform, 2025.

[JKVdS+21] Shawn M. Jones, Martin Klein, Herbert Van de Sompel, Michael L. Nelson, and Michele C. Weigle. Interoperability for accessing versions of web resources with the Memento protocol. In *The web as history*. Springer, 2021.

[KAA+17] Mat Kelly, Lulwah M. Alkwai, Sawood Alam, Michael L. Nelson, Michele C. Weigle, and Herbert Van de Sompel. Impact of URI canonicalization on memento count. In *Proceedings of the 17th ACM/IEEE joint conference on digital libraries*, 2017.

[KL20] Daphne Keller and Paddy Leerssen. Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation. In Joshua A. Tucker and Nathaniel Persily, editors, *Social Media and Democracy*, SSRC Anxieties of Democracy, pages 220–251. Cambridge University Press, Cambridge, 2020.

[KSBVdS19] Martin Klein, Harihar Shankar, Lyudmila Balakireva, and Herbert Van de Sompel. The memento tracer framework: Balancing quality and scalability for web archiving. In *Digital libraries for open knowledge: 23rd international conference on theory and practice of digital libraries, TPDL 2019, oslo, norway, september 9–12, 2019, proceedings*, volume 11799 of *Lecture notes in computer science*, pages 163–176. Springer, Cham, 2019. arXiv: 1909.04404.

[KVdSS+14] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. Scholarly context not found: One in five articles suffers from reference rot. *PLOS ONE*, 9(12):e115253, 2014.

[Lat05] Bruno Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, July 2005.

[Mak21] Ruwadzano Patience Makumbe. Sam Dubberley, Alexa Koenig and Daragh Murray (editors), Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation and Accountability. *Journal of International Criminal Justice*, 19(1):229–233, March 2021.

[MMK22] Daragh Murray, Rachel McDermott, and Alexa Koenig. Mapping the use of open source research in UN human rights investigations. *Journal of Human Rights Practice*, 14(2):554–578, 2022.

[MS95] Salvatore T. March and Gerald F. Smith. Design and natural science research on information technology. *Decision Support Systems*, 15(3):251–266, 1995.

[MSM15] Jim Maddock, Kate Starbird, and Robert Mason. Using historical twitter data for research: Ethical challenges of tweet deletions, 2015.

[Nat19] National Digital Stewardship Alliance. Levels of digital preservation, version 2.0, 2019.

[NR21] Julian Nida-Rümelin. Digital humanism and the limits of artificial intelligence. In H. Werthner, E. Prem, E. A. Lee, and C. Ghezzi, editors, *Perspectives on Digital Humanism*. Springer, 2021.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

[NRW22] Julian Nida-Rümelin and Nathalie Weidenfeld. *Digital Humanism: For a Humane Transformation of Democracy, Economy and Culture in the Digital Age*. Springer International Publishing, Cham, 2022.

[Pew24] Pew Research Center. When online content disappears: Link rot and digital decay on government, news and other webpages, May 2024.

[PRE15] PREMIS Editorial Committee. PREMIS data dictionary for preservation metadata, version 3.0. manual, Library of Congress, 2015. Version number: 3.0.

[PTRC07] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3):45–77, December 2007.

[RA18] Benjamín Reyes Ayala. *A grounded theory of information quality in web archives*. phd, University of North Texas, 2018.

[Ris19] Sandra Ristovska. Human rights collectives as visual experts: the case of Syrian Archive. *Visual Communication*, 2019.

[Rol14] Gregory Rolan. Towards Archive 2.0: issues in archival systems interoperability. *Archives and Manuscripts*, 42(3):257–271, 2014.

[RSM22] Maud Reveilhac, Stephanie Steinmetz, and Davide Morselli. A systematic literature review of how and whether social media data can complement traditional survey data to study public opinion. *Multimedia Tools and Applications*, 81(7):10107–10142, March 2022.

[RT23] Iavor Rangelov and Ruti Teitel. The justice archive: transitional justice and digital memory. *London Review of International Law*, 11(1):83–109, March 2023.

[SBG04] Detmar Straub, Marie-claude Boudreau, and David Gefen. Validation Guidelines for IS Positivist Research. *Communications of the Association for Information Systems*, 3, January 2004.

[Sny19] Hannah Snyder. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104:333–339, November 2019.

[Ste46] S. S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

[STvO+18] Thaer Samar, Myriam C. Traub, Jacco van Ossenbruggen, Lynda Hardman, and Arjen P. de Vries. Quantifying retrieval bias in Web archive search. *International Journal on Digital Libraries*, 19(1):57–75, March 2018.

[SWQY19] Nicolas P. Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication*, 13:18–18, March 2019.

[Syr25] Syrian Memory Organization. Syrian Memory: Digital Archives and Documentation of the Syrian Conflict, 2025.

[Tar24] Stefan Tarnowski. The Beginning of an Idea about an End: On Digital, Diasporic, Syrian Archives. *Visual Anthropology*, 37(1):35–55, January 2024.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

[TFC25] Amaury Trujillo, Tiziano Fagni, and Stefano Cresci. The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media. *Proc. ACM Hum.-Comput. Interact.*, 9(2):CSCW187:1–CSCW187:28, May 2025.

[UU20] UN Office of the High Commissioner for Human Rights and UC Berkeley Human Rights Center. Berkeley protocol on digital open source investigations: a practical guide on the effective use of digital open source information in investigating violations of international criminal, human rights and humanitarian law. Technical report, United Nations, New York and Geneva, 2020.

[VdSNS13] Herbert Van de Sompel, Michael L. Nelson, and Robert Sanderson. HTTP framework for time-based access to resource states — Memento (RFC 7089). Technical report, Internet Engineering Task Force, 2013.

[Vic23] Paula Vicente. Sampling Twitter users for social science research: evidence from a systematic review of the literature. *Quality & Quantity*, 57(6):5449–5489, December 2023.

[Web23] Webrecorder. Web archive collection zipped (WACZ) format specification, 2023.

[Web25] Webrecorder. Browsertrix documentation, 2025.

[Wer23] H. Werthner. *Introduction to Digital Humanism*. Springer, Cham, 2023.

[WH00] Rüdiger Wirth and Jochen Hipp. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–40, Manchester, UK, 2000.

[WH25] Qi Wang and Andrew Hoskins, editors. *The Remaking of Memory in the Age of the Internet and Social Media*. Oxford University Press, January 2025.

[Whi24] Elizabeth White. Closing cases with open-source: Facilitating the use of user-generated open-source evidence in international criminal investigations through the creation of a standing investigative mechanism. *Leiden Journal of International Law*, 37(1):228–250, March 2024.

[Wie14] Roel Wieringa. *Design science methodology for information systems and software engineering*. Springer, Berlin, Heidelberg, 2014.

[WLA+19] Hannes Werthner, Edward A. Lee, Hans Akkermans, Moshe Vardi, Carlo Ghezzi, Nadia Magnenat-Thalmann, Helga Nowotny, Lynda Hardman, Oliviero Stock, James Larus, Marco Aiello, Enrico Nardelli, Michael Stampfer, Christopher Frauenberger, Magdalena Ortiz, Peter Reichl, Viola Schiaffonati, Christos Tsigkanos, and William Aspray. Vienna manifesto on digital humanism, 2019.

[WNAG23] Michele C. Weigle, Michael L. Nelson, Sawood Alam, and Mark Graham. Right HTML, wrong JSON: Challenges in replaying archived webpages built with client-side rendering, 2023.

[WPLG22] Hannes Werthner, Erich Prem, Edward A. Lee, and Carlo Ghezzi, editors. *Perspectives on Digital Humanism*. Springer International Publishing, Cham, 2022.

[You26a] YouTube Help. Age-restricted content, 2026.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[You26b] YouTube Help. Change video privacy settings, 2026.

[You26c] YouTube Help. Channel or account terminations, 2026.

[You26d] YouTube Help. Videos locked as private, 2026.

[Zub18] Arkaitz Zubiaga. A Longitudinal Assessment of the Persistence of Twitter Datasets. *JASIST*, 2018.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.