Internet Histories
Digital Technology, Culture and Society

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rint20

Genealogy of an archive. The birth, construction, and development of the World Wide Web collection at CERN

Martin Fomasi, Deborah Barcella, Eleonora Benecchi & Gabriele Balbi

To cite this article: Martin Fomasi, Deborah Barcella, Eleonora Benecchi & Gabriele Balbi (2023) Genealogy of an archive. The birth, construction, and development of the World Wide Web collection at CERN, Internet Histories, 7:3, 277-294, DOI: 10.1080/24701475.2023.2238254

To link to this article: https://doi.org/10.1080/24701475.2023.2238254

Published online: 10 Aug 2023.

INTERNET HISTORIES
2023, VOL. 7, NO. 3, 277–294
https://doi.org/10.1080/24701475.2023.2238254

Routledge
Taylor & Francis Group

Genealogy of an archive. The birth, construction, and development of the World Wide Web collection at CERN

Martin Fomasi, Deborah Barcella, Eleonora Benecchi and Gabriele Balbi

Institute of Media and Journalism, Università della Svizzera Italiana, Lugano, Switzerland

ABSTRACT

Web and Internet historians have never been able to consider the sources preserved at CERN because of a 30-year closure law. The WWW collection is of major importance not only because it is located where the Web was born but, more importantly, because it preserves documents produced during the early and little-known stages of its development. Our study has a qualitative approach and is based on in-person discussions, e-mail exchanges, and a focus group we conducted with five main actors responsible for the birth and development of the WWW collection at CERN. Through this method, we co-constructed with them a discourse, which we later analysed through inductive thematic analysis. We extracted six main topics reflecting the principal themes represented in the collection: reasons for creating a specific collection of web-related documents; salient moments in the history of the collection; discussion about its naming; issues about the originality of the documents; and future digitisation projects. This paper may be of interest to web historians and archivists looking for an overview and hidden reasons for the creation of the collection.

ARTICLE HISTORY
Received 18 January 2023
Revised 11 July 2023
Accepted 12 July 2023

KEYWORDS
Web history; CERN archive; WWW collection

1. Introduction

The World Wide Web (W3 or WWW) was conceived at CERN, the European Organisation for Nuclear Research, from 1989 until it was released into the public domain in 1993. As a result, CERN preserves the most relevant documents about the origins of the Web, which are currently available after the 30-year rule of archive closure.

This article focuses on the creation of the WWW collection within the CERN Archive, the place where most of the documents on the Web should be preserved. It offers a genealogy of the archive, investigating its origins, the main actors involved in its creation, the reasons behind its establishment, and the significant issues that have emerged over time. The study of the origins of the archive clarifies why this collection can't be considered a comprehensive repository of Web sources, as it reveals the

CONTACT Martin Fomasi martin.fomasi@usi.ch Institute of Media and Journalism, Università della Svizzera Italiana, Via Buffi 13, Lugano 6900, Switzerland

existence of other Web documents preserved outside CERN. Those archives aren't included in this paper, but just mentioned.

This critical analysis is relevant because, even if previous papers explored the CERN archive in general (Hollier, 2005, 2007, 2011), no research has been devoted to the WWW collection. Moreover, all previous papers on the CERN archives were written by people who worked within the same institution, making this the first study on a CERN archival collection conducted by researchers who didn't work at CERN.

The paper is structured as follows: after having provided a theoretical framework and a methodological overview of the present study, it considers CERN's WWW collection by reviewing the information available about the origin and development of this archive and the results of our research focusing on its creation and evolution. This research is based on the analysis of discourse on such archives co-constructed with CERN archivists and relevant figures in the history of the Web through in-person discussions, e-mail exchanges, and a focus group. In the conclusions, we weave together the different data collected with the theoretical framework to provide a critical analysis of our object of study.

2. Theoretical framework

The "classic" STS authors Bowker and Star (1999) invited researchers to reconsider the implementation of classification and other related fields such as the development and deployment of standards and archives. They stressed the importance of acknowledging the "ethical and political value modulated by local administrative procedures" that shape information systems, making them both "active creators of categories as well as simulators of existing categories" (p. 321).

By adopting Bowker and Stars theoretical framework, we aim to study "spaces and flexibility that are otherwise lost forever" (p. 321) and the moral and governmental principles that influenced the formation of this (and other) archive and its categories. This should allow re-emerging specific choices made over time and then simply forgotten, as it is common in any classification systems set up in organisations like CERN. Therefore, nearly 30 years after the Web release in the public domain, this analysis investigates the negotiations between actors inside CERN and events leading to the creation of the WWW collection. In this sense, we show that the responsibility of the preservation action is shared among multiple actors (Latour, 1994, 1999). As this research considers not only the history of archive development but also his classification system, we sought to identify actors and documents excluded in the development process.

The WWW collection at CERN will be treated under the notion of "total archive" described by Jardin and Drage (2018). A total archive is more than just a collection of documents: it involves a complex assemblage of people, organisations, technologies, and practices that work together to preserve and provide access to information. The WWW collection could be seen as an example of a total archive because it is shaped by the interactions of different actors, including CERN staff, researchers, software developers, and archivists, involved in its creation. The collection, as any other total archive, is the result of a process occurring in a large-scale social system that emerges through the combination of specific social settings and groups that fix and frame a

particular moment in an object's history. In other words, total archives reconstruct the world according to the image of their guiding logic implemented during their creation process (Jardine & Drage, 2018). Therefore, understanding the genealogy of an archive is critical in highlighting the rationale behind its creation, which can provide historians and archivists with a tool for archival appraisal (Cook, 2011).

By adopting the above theoretical frameworks, we gain a more nuanced understanding of the complex dynamics that shaped the creation and preservation of this historically significant archive.

3. Methodology

After revising secondary literature on the history of the CERN archives, we focused on the WWW Collection, which shares some of the characteristics of CERN's general archive. We adopted a qualitative approach that involved co-constructing a discourse around the archive through a reflexive practice that critically examines "the process, context, and outcomes of research and interrogates the construction of knowledge" (Finlay, 2012, p. 317). This was pursued by (1) comparing, through in-person discussions and e-mail exchanges, our collected data, and running hypotheses with the recollections of people who were involved in the WWW Collection; (2) using the information thus collected, we identified a set of relevant issues and people that could be used as a basis for exploring specific themes through a focus group; and 3) following previous examples (Fratila & Sionis, 2006; Potter & Hepburn, 2012; Roulston, 2016; Wertz, 2011), we devised an approach where both researchers and interviewers were invited to reflect on and consider their part in the research process. CERN people involved in our research have been:

- Robert Cailliau: Tim Berners-Lee's main collaborator in the development of the WWW.
- James Gillies: Former head of the CERN communications group and former official CERN spokesperson. Co-author of *How the Web was Born: the story of the World Wide Web* (Gillies & Cailliau, 2000).
- Jens Vigen: Deputy leader of the Scientific Information Service and section leader of the CERN archive.
- Sandrine Reyes: Archivist-assistant at CERN.
- Anita Hollier, CERN archives manager, was also contacted and answered questions dealing with the general archive but did not take part in the focus group.

Our first contact with these people were through a series of e-mail exchanges followed by in-person meetings at CERN. As both a summary of relevant topics emerged and a follow-up of the first e-mail exchanges and in-person meetings, we sent the same e-mail to the participants asking the following questions: (a) Are you aware of secondary sources on the building of the WWW collection at CERN?; (b) Who were the actors involved in the building of the WWW collection at CERN?; (c) Who has kept the documents safe over time? (We found a document from 1997 about the policies that were in place at that time, which we are attaching); (d) What were and are the reasons for having a special collection on the Web?; (e) Selection

criteria: how were the documents selected to be part of the collection?; (f) Web-related documents that could be included in other collections. Which is the policy?; (g) When was the politics of document digitalisation and open access developed? Has it changed over time?; (h) What are the criteria for adding new historical documents to the WWW collection that could be added in the future?

Based on the answers, we produced assertive contributions during the exchanges with the CERN people only in case of discrepancy between data available within the WWW collection or previous statements on the different topics, and the participants' contribution. Already in this preliminary phase, we can thus observe a co-construction of discourse since the researchers actively contributed to the memory-activating activities and reflected on their role and experience in exploring the research topic with the research participants.

The results of this phase provided new data that helped us develop a guide comprising more specific topics of interest to be deployed within a focus group setting intended as a follow-up phase that pursued exploratory aspects of the analysis (Wilkinson, 1998). Following Vaughn et al. (1996), the focus group was organised to elicit participants' feelings, attitudes, and perceptions of a specific topic and conducted by a trained moderator, the research leader, and an expert in digital media history. In agreement with Morgan (1998, p. 33), this focus group involved research on a specific topic, the moderator kept the participants focused, and the participants were invited to interact with each other.

To allow for contributions from the research team that was present but silent during the focus group, visual stimuli were prepared to elicit participants' recollection of the research topic but also to challenge eventual discrepancies with available data. Specifically, the stimuli were based on materials gathered within the WWW collection or on previous contributions from the participants collected by the team during the first phase of the study. Additionally, the participants were invited to say as much as they could remember about their experience with the WWW collection with a pre-designed list of topics that were delivered in the form of open questions (Puchta & Potter, 2004).

The session was video-recorded and transcribed in a way that "captures action" (Potter & Hepburn, 2012, p. 559) thus including significant features such as pauses, overlaps, and various non-linguistic features. Under the "researcher reflexivity" stance and to follow the model of the co-construction of discourse (Potter & Hepburn, 2012; Speer & Stokoe, 2014; Wertz, 2011), the participants were allowed to review the focus group transcription and the first draft of the authors' analyses of their contributions. This allowed us not only to gain informed consent to use the focus group data but also to check the participant's accounts and descriptions of their lived experiences in terms of truthfulness and accuracy. Additionally, this model also offered the researchers the opportunity to turn back to examine the resources used in the generation of research data and thus their role in co-constructing the participants' account of the research topic.

To analyse the focus group, we used a reflexive thematic analysis inspired by Braun and Clarke (2022). The first step of the analysis—the "familiarisation with the dataset" phase—consisted of familiarising ourselves with the collected data by individually reading the e-mail exchange and the focus group transcript several times. In the

second phase—the "coding" phase—we proceeded by identifying text segments that provided answers to our questions and discussing their characteristics together. Based on what we talked about, we came up with six tentative topics and talked about how important they were until all the researchers agreed. This is the same as the "generating initial themes" phase. We developed and reviewed themes by associating them with smaller portions of the text so that the meanings associated with each theme would emerge. Based on the portions of the text selected, we renamed certain categories and defined what they stand for. Six themes were thus constructed and defined: those are discussed in the sections 4 to 9.

4. Who

This section aims to show who were the main actors in the collection building. The WWW collection (CERN, n.a.–a) is a group of documents that were put in the CERN archive in 2001, 2013, 2021, and 2022 by five people, with Robert Cailliau playing a key role, having given to CERN the greatest part of the collection. The section 'Scope and Content' of the CERN Archive Guide provides the following overview of the collection:

> The collection covers the period 1988–1999 and includes correspondence, reports, notes, minutes of meetings, administrative documents, and conference and other presentations. It deals with the development of the WWW and efforts to secure support and funding for it. (CERN, n.a.–b)

Figure 1 makes visible one big absentee: Tim Berners-Lee, the person who has always been considered the inventor of the Web. The figure also shows archival materials were not added to the WWW collection in a straight line. Instead, the folders were made in a way that changed over time.

Robert Cailliau is considered Tim Berners-Lee's main collaborator. They co-wrote the proposal that CERN accepted in 1990 (Gillies & Cailliau, 2000), following 1989 Tim Berners-Lee's individual submission, which was held pending by Mike Sendall. The proposal's central goal remained the same: to facilitate information retrieval for scholars and scientists visiting CERN for short periods, as underscored by the frequent use of the term "information" (Bory et al., 2016). In 1990, Robert Cailliau founded CERN's Web Office and, in December 1993, launched the World Wide Web Conference series (Gillies & Cailliau, 2000). After Tim Berners-Lee left CERN to lead the World Wide Web Consortium (W3C) from the Massachusetts Institute of Technology (MIT), Robert Cailliau helped transfer web development from CERN to the W3C. The WWW collection is composed mainly of the documents he collected (CERN-ARCH-WWW-1-*, CERN-ARCH-WWW-2-*, CERN-ARCH-WWW-3-*).

Mike Sendall, leader of the Online Computing group, the CERN section where Tim Berners-Lee worked under the supervision of Peggie Rimmer (Gillies & Cailliau, 2000) , is the second relevant actor appearing in Figure 1. When Tim Berners-Lee submitted the first web proposal, Mike Sendall wrote the note "Vague but exciting" without rejecting or accepting it. He later encouraged him to develop the first working prototype of a Web browser (Gillies & Cailliau, 2000). During the focus group, it emerged that when Mike Sendall passed away, Robert Cailliau kept more of his documents that were added to the WWW collection (CERN-ARCH-WWW-4-*).

CDS	Files	Content	Received from	Period	Tot records
CERN-ARCH-WWW-1-*	Files of Robert Cailliau	Main series – development of the World Wide Web	Received from Robert Cailliau, 15 February 2001 (CERN-ETT Division)	1988/1999	31
CERN-ARCH-WWW-2-*		Subject files on the World Wide Web		1986/1997	25
CERN-ARCH-WWW-3-*		First World Wide Web Conference 1994		1993/1994	9
CERN-ARCH-WWW-4-*	Files of Mike Sendall	World Wide Web		1989-1997	10
CERN-ARCH-WWW-5-*	Peter Jurcso	Copy of NeXT computer hard disk	Received from Dan Noyes 27 May 2013 CERN-DGU	1999	1
CERN-ARCH-WWW-6-*	Files of Ben Segal Info	Remote Procedure Call (RPC)	Received from Ben Segal 22 July 2021	1983/1986	1
CERN-ARCH-WWW-7-*	Files of James Gillies	WEB Book – "How the Web was Born, the story of the World Wide Web"	Received from James Gillies 8 February 2022 CERN-IR-ECO	1985/1999	22

Figure 1. Adaptation of a summary table of the WWW collection (Jens Vigen, personal communication, June 2, 2022).

The third person in Figure 1 is Peter Jurcso, a CERN systems analyst who made a backup of Tim Berners-Lee's NeXT hard drive, a computer used by Tim Berners-Lee for writing the Web client, Web server, and the client program WorldWideWeb browser. Peter Jursco gave the backup to Dan Noyes, head of the Content section in the CERN communications group until January 2016 (CERN, n.a.–c) and major contributor of the first website restoration project (Noyes, 2013). Now, the copy of the NeXT computer hard disk is in the WWW collection (CERN-ARCH-WWW-5-*).

The fourth person is Ben Segal, who introduced the Internet at CERN and helped Tim Berners-Lee convince Mike Sendall to develop Remote Procedure Calls (RPC) to control the data acquisition systems of the Large Electron-Positron Collider (LEP). Its material goes under the code CERN-ARCH-WWW-6-*.

The latest person is James Gillies, who headed CERN's communications group until January 2016 and now works in the Education, Communication and Outreach group (CERN, n.a.–d). James Gillies co-authored with Robert Cailliau the book *How the Web*

Was Born, drawing on numerous interviews and documents. The reference code CERN-ARCH-WWW-7-* for James Gillies refers to the material he gathered and used for the book *How the Web Was Born*.

During the focus group, Robert Cailliau also emphasised several times the role played by Anita Hollier, who doesn't appear in Figure 1. Anita decided to make room in the CERN archives for the materials and make a catalogue of them. She also created the above-mentioned descriptive of the collection in the CERN archive catalogue which she kept updated until February 2022 (CERN, n.a.–b). Despite our attempts, it wasn't possible to assess the degree of autonomy Anita had in taking this decision.

The documents are stored in 99 folders covering approximately 5 linear meters. During a workshop we had with our research partners, including some members of CERN, Anita Hollier declared that the collection spans the years 1988 to 1999 and consists of mail, reports, minutes of meetings, administrative documents, conference materials, and other presentations. Archival documents talk about many things related to the web, such as proposals, policy discussions (like how software should be distributed and released into the public domain), contracts between CERN and other organisations like INRIA and the European Union, documents about the creation of the W3C, awards for the best technology of the year, and many other things.

The documents in the WWW Collection are gathered in records and catalogued following a reference code that reflects a specific topic and the person who preserved the document. For example, the reference code CERN-ARCH-WWW-1 indicates Robert Cailliau's files, and it includes documents referring to the topic "Main series development of the Web". Each record relating to CERN-ARCH-WWW-1 is then indicated by a different three-digit number following the reference code (e.g. CERN-ARCH-WW-1-001). The reference codes don't necessarily reflect chronological order, and no single reference code exists for each document contained in a record.

5. When

The second theme focuses on time and, specifically, when the collection was built. During the data collection process, we tried to understand the precise establishment of the WWW collection. Robert Cailliau was the first person to save documents for the long term, even before CERN made the Web available to the public. During the focus group, he claimed: "I started keeping bits and pieces from about 1992 onwards, including such trivia as conference badges (which I unfortunately no longer have), T-shirts, pins, and other memorabilia".

As emerged through the written exchanges with James Gillies and Robert Cailliau the editing of the book was an important event for the growth of the collection. According to James Gillies and Robert Cailliau, the work started around 1999. Figure 1 suggests that Robert Cailliau handed over the collected material to Anita Hollier in 2001. By contrast, the material collected specifically for the book by James Gillies was stored at his home and delivered to Sandrine Reyes in December 2022.Thus, the WWW collection started to grow in 2000, but according to people involved, it isn't possible to say when it started. In the focus group, Robert Cailliau remarked: "[…]

you're trying to look for a sort of single moment, of a single Big Bang or something like that. But it didn't really happen like that".

Establishing a precise date for the creation of the collection isn't only impossible but also not relevant as expected. Rather, the WWW collection is marked by some noteworthy events that should be registered: the year Robert Cailliau began collecting the material (1992), some indication of when the collection grew rapidly (about 1999), the time frame of the catalogue redaction, and the storage of material in the archive in distant periods (2001, 2013, 2021–2022).

6. Why

Our research shows that the rationale behind the birth of the WWW collection is connected to three key drivers which varied in their significance and interdependence. First, a driver that we labelled as "publication driver", namely the desire to document the history of the Web in the form of a book—*How the Web Was Born: The Story of the World Wide Web* (Gillies & Cailliau, 2000)—had a significant impact on the establishment of the archive. Second, the driver of "historical awareness", or the belief and forecast that the Web would become prominent in the future as well as its historical sources, had an important impact on the "circumstantial" driver and pushed to start to persevere material relatively early after the first proposal. Third, the creation of the archive was also the result of a "personal commitment" driver, resulting in the private effort of a small number of individuals.

The "publication driver" is the most important in triggering the creation of the archive and it brought an expansion of the actors involved. Everything started when Robert Cailliau's wife, Susan (former Divisional Records Officer of a CERN division), convinced him a book should be written on the early history of Web technology. The preparation of the book was the principal factor that prompted Robert Cailliau and others to keep and collect material on the WWW and to organise these files into a system. In the focus group, Robert Cailliau claimed:

> […] to say that story in a few sentences, my dear wife […] also suggests writing it all down. I thought we had to get [in touch with] somebody who was not involved with CERN and was not involved with the technology and so forth to write this book. And I contacted several [people] and they all said "no" or declined or found it not interesting. And finally, James [Gillies] came along and said, "Well, we'll have to do it ourselves after all". So then […] I collected some of the documents [that are now in the WWW collection.

And it was during the book-writing period that James Gillies, CERN's communications manager, came on the scene to collect additional material that would later be deposited in the WWW collection. During the focus group, James Gillies stated:

> The book was written in 1999. It was basically a one-year project. […] I remember going and looking at computing newsletters and talking to people. Everybody you know, I interviewed a lot of people, and they made things [Web-related records] available to me.

The drive of "historical awareness" is again triggered by Susan Cailliau. During the focus group, Robert Cailliau remarked that his wife was aware of the importance of

the Web, so much so that one day in the early 90s, she said to him: "[The Web] is going to be extremely important, and you should start organising it and keeping your documents […] because this is going to be very important".

The most surprising aspect emerging from the drivers is that, despite perceiving the Web as a successful project, its creators didn't place much importance on preserving sources for future historical studies. In the focus group, Robert Cailliau candidly admitted that:

> I had just not thought of it [collecting and preserving historical material]. I think the historical aspect was neglected until [my wife] said that and for Tim [Berners-Lee] came even later, and for some of the other guys, I think it never came.

This suggests that, in the early years of the Web, people who were connected but not directly involved in the Web development showed more awareness of its historical significance than those who created it.

Jens Vigen emphasises that "historical awareness" influenced the personal and institutional collection as well as the book. During the focus group, he stated that without it, neither the book nor the WWW collection would have been created:

> But I would guess that all these documents were collected in the process of writing the book, but probably deposited by Robert a couple of years later. […] I mean it's a bit of speculation, but I believe that what you refer to now as the collection was probably built up as Robert became aware that preserving history is important. And then James used this material when he was writing. And then, in the end, it was deposited in the archive.

When it comes to the third driver, "personal commitment", Susan Cailliau was the one who started urging her husband to preserve all documents and objects related to the Web development: "At one point at the beginning of the 90s […] my dear wife [pushed] me to do historic stuff [so I started] […] keeping documents and filing them and keeping every little bit like the T-shirts". Thus, Robert Cailliau and his wife preserved the technology's records privately and without CERN's pressure. Therefore, the majority of the WWW collection is the result of Robert Cailliau's personal and material contribution to the development and promotion of the technology.

In summary, the three drivers collectively stimulated the inception of the WWW collection, with each driver reinforcing the others. Ultimately, the archive's creation was made feasible by the dedication and the interests of a few individuals who aimed to write a book on the history of the Web, recognised its historical significance and diligently collected records. This effort culminated in a vast collection of material that spanned five-meter shelves.

7. Naming

So far, we have used the term "WWW collection" to refer to the specific and limited set of documents concerning the development of the web preserved at CERN, but the official name of this collection is uncertain. First, the name appears differently depending on the documents consulted. Second, the use of the term "collection"

created a controversy during the e-mail exchanges. In an e-mail sent to us in June 2022, Robert Cailliau wrote:

> Perhaps Anita and Sandrine can help to say exactly what is meant (officially) by WWW collection. There are at CERN a few official collections I know of but I'm not sure at all that there is an official "WWW collection". (R. Cailliau, personal communication, June 1, 2022)

To clarify this point, before the focus group, we considered two CERN websites: the CERN Scientific Information Service and the CERN Document Server. On the first one, we can find the title "World Wide Web" and the term "collection" to describe the section "Scope and Content" (see section 4 of this paper). On the second, as shown in Figure 2, the collection research bar of the CERN Document Server provides the title "the World Wide Web, WWW (Archives)". Inside the same research bar, we can also find a collection titled "World Wide Web". However, this latter is composed only by 6 records. A copy of the document releasing the Web in the public domain, the others are mainly photos dated after 2007 (CERN, n.a.–e).

This difference could be explained by the fact that the CERN Scientific Information Service webpage is still a draft that must be completed and harmonised with all the other information services. For practical reasons, we decided to use the "WWW collection" naming.

Nevertheless, this ambiguity and multiple labelling provided a topic of discussion for the focus group. As we showed above, the term "collection" appears on both CERN's websites. We showed Figure 2 to the participants and ask them during the focus group to discuss the use of the term "collection". Robert Cailliau opened the talk by evoking the dimension of wholeness:

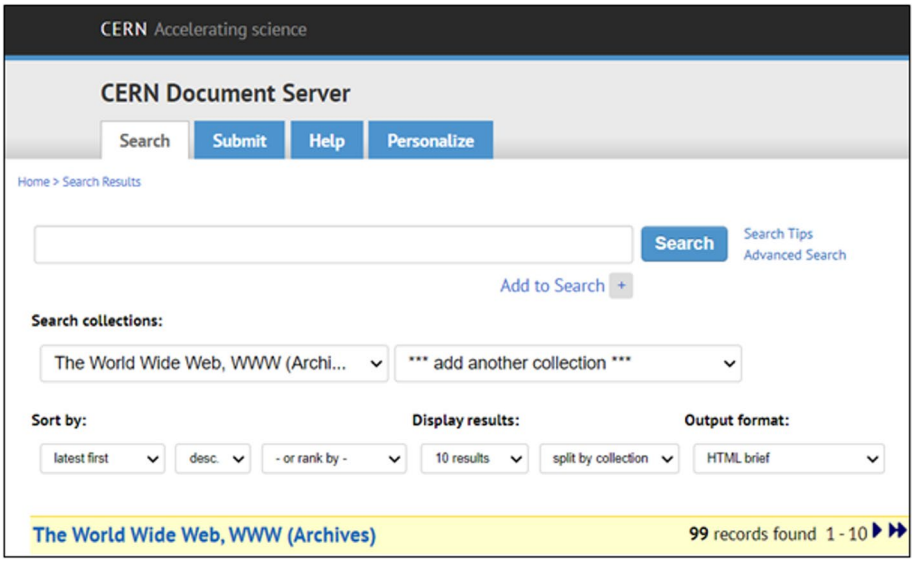

Figure 2. CERN Document Server website showing the World Wide Web, WWW (archives) in the collections research bar (last visit: 17 April 2023).

> I think there are documents relating to the web, but they are not tagged. Jens, correct me if I'm saying something wrong, but I don't think there is some sort of a special tag that says. This document is not only personal and important for the archives of Carlo Rubbia [former CERN director general], but it also refers to the World Wide Web.

As indicated by this quote, participants had to co-construct a discourse indicating that the term "collection" doesn't refer to a complete set of all the documents relating to the web preserved at CERN. The incompleteness of the WWW collection created a controversy around the abstract matter of its naming and specifically around the meaning to be associated with the term "collection" itself.

8. What's missing in the WWW collection?

The fifth theme we isolated during the analysis deals with what is missing in the WWW collection. The fact that the WWW collection is a partial storage was the subject of a preponderant conversation during the focus group, as the collection gathers only some of the many documents created in the early years of the Web. According to the participants in the meeting, the other documents can be grouped in three places: (1) other collections still within the CERN archive; (2) in private or institutional collections outside the CERN archive; and (3) in unknown locations and therefore here labelled as "disappeared."

8.1. Inside CERN

Focus group participants agree that many documents of the Web are in other collections of the CERN archive: "[The WWW collection] is very incomplete because […] at least Pier Giorgio [Innocenti]'s documents are not there" although Robert Cailliau claimed that: "I have had many sessions with Pier-Giorgio [Innocenti], to which he brought several binders full of important documents from his personal collections" (R. Cailliau personal communication, November 15, 2022).

Pier Giorgio Innocenti was one of Robert Cailliau's superiors and a key player in the establishment of the W3C (Gillies & Cailliau, 2000). During his tenure at CERN, he also played a crucial role in negotiating an orderly development of the Web protocol between CERN, the European Commission, MIT/LCS, and INRIA. Robert Cailliau and Pier Giorgio Innocenti engaged in a close communication exchange for this reason. At CERN archives, there is an entire collection belonging to Pier Giorgio Innocenti and called "CERN-ARCH-INNOCENTI," The number record 004 includes documents related to the 1994–1995 negotiation between CERN, the European Commission, MIT/LCS, and INRIA, to safeguard the development of the Web, especially the development of its protocol. Indeed, Sandrine Reyes declared that if a Web-related source belongs to a CERN member's collection, archivists can't make a copy to put it in the Web collection to have a duplicate.

By performing an advanced search on the CERN document server, we can see that some Web documents may also be in the CERN collection of the Data Handling Division (DD), where many David Owen Williams's documents are stored. David Owen Williams joined CERN's Data Handling Division (DD) in 1966. He was a big supporter

of the Internet's growth in Europe, not just as a tool for science but also to help the economy grow in Europe as a whole (De Gregorio & Manai, 2014). For this reason, he also supported Tim Berners-Lee and Robert Cailliau in the development of the Web.

Other papers are also found in the collections of the Proton Synchrotron Division (PS), where Tim Berners-Lee and Robert Cailliau worked in 1980; the Electronics and Computing for Physics Division (ECP), another division where Robert Cailliau worked; and the Computer Network Division (CN), where Tim Berners-Lee worked; and in the collection of the Computer Newsletter, which was used to update users of the CERN computer facilities on developments and trends in the area (CERN, n.d.–f).

8.2. Outside CERN

Certain records aren't just outside the WWW collection but are also outside CERN, even though Circular Letter No. 3 of 1997 declared that "CERN is responsible for its documents and files and their preservation." (p. 1). Many valuable historical documents on the Web may be found in the private collections of individuals who worked on the Web project at CERN, as well as in other institutions' archives.

Most missing records are Tim Berners-Lee's ones. Robert Cailliau told us in a 2022 personal communication that Tim Berners-Lee wasn't accustomed to writing or keeping many documents. Robert Cailliau can't recall any printout-containing folders in his office. According to him, the absence is thus more akin to nonexistence. Nonetheless, research revealed that some of his papers could be preserved in the MIT archives. The first two MIT archive collections on Tim Berners-Lee and the World Wide Web are made up of records made by Albert Vezza, known for his role in the establishment of the W3 Consortium (Gillies & Cailliau, 2000, p. 331), and the Laboratory for Computer Science (MIT, n.d.–a, n.d.–b). Some of these records are thesis proposals, computer tapes, reports, memos, abstracts, notes, books, transcripts, slides, transparencies, and notes from depositions. There is also a collection entirely dedicated to the World Wide Web Consortium (1994), including original papers and other materials from the Laboratory for Computer Science (LCS) (MIT, n.d.–c) Lastly, there is a recent collection potentially containing Tim Berners-Lee's documents (from 2007 to 2011) created by MIT President Susan Hockfield and includes letters, lectures, MIT centers, labs, programs, schools, companies, the Executive Committee, and the Visiting Committee (MIT, n.d.–d).

CERN decided around the end of 1994 to devote all its resources to the development of the Large Hadron Collider (LHC). As a result, CERN declined the offer to host the W3C in Europe. In April 1995, MIT chose INRIA as their partner (INRIA, n.d.). This suggests INRIA as another possible location of documents on the Web. Another place that has hosted the Web Consortium (W3C) and may have some documents is Keio University in Japan.

Other Web documents that aren't currently at CERN are Mike Sendall's papers. During the focus group, Robert Cailliau affirmed that: "I don't know where all of Mike Sendall's documents are. I think a lot of them are in his house with his wife [Peggy Rimmer]." In 2022, Peggy Rimmer decided to donate these documents to the British Library in the UK and negotiations are currently underway also with CERN archives.

When this paper was written different options were on the table, but probably in a near future other crucial documents on the origins of the Web at CERN will be available for researchers.

Focus group respondents pointed out that, according to them, other Web-related documents may be in the personal archives of Gottfried Keller (Robert Cailliau's superior who has probably conserved a copy of an information system development proposal Cailliau wrote before Tim Berners's Lee first proposal for the Web), Jean-François Abramatic (Director of the Development at INRIA and one of the chairmen of the W3C) (Gillies & Cailliau, 2000), François Flückiger (CERN computer scientist who helped the management of the Web after Tim Berners-Lee joined the MIT) (Gillies & Cailliau, 2000)), and Mark Weber (the curator at the Computer History Museum in California) (R. Cailliau, personal communication, November 15, 2022). This would increase the percentage of Web-related documents produced by people who worked at CERN that isn't in CERN's possession. Robert Cailliau also said that he has some documents at home, but, according to him, these are only copies of documents already in the WWW collection, so nothing new compared to what can already be found there. During the focus group, Jens Vigen informed us that it is difficult for the organisation to claim ownership since it has "No resources to sit down and try to figure out which could be the potential document holders and obtain it from them". So, CERN doesn't have any plans right now for how to find important documents for the history of the Web. Furthermore, the CERN Circular Letter related to archive material didn't come out until 1997, so in the early years of the Web, CERN's need for document retention wasn't made explicit. In the specific case of Peggie Rimmer, we see instead a trust issue with the CERN archive. During the focus group, James Gillies emphasised that in her opinion the British Library is a safer repository than CERN Archives.

The fear of losing personal material is a factor that increases distrust of CERN and amplifies the phenomenon of external collections. James Gillies thinks that the lack of trust in CERN may be due to the way that documents are stored. For example, CERN members have been asked to throw away administrative documents that were deemed unimportant more than once. The problem is that the person in charge of deciding which documents to store used his or her judgment. During the focus group, James Gillies claimed that this disorganisation:

> […] might also explain, to a certain extent, Peggie's wish to give what she has to the British Library rather than to CERN. Also, in the messy situation that Robert was just describing, Peggie must have lived through that too. So, she's aware of that as well.

8.3. The disappeared

Besides having Web documents misplaced or outside the CERN archive, some fundamental historical sources have disappeared:

> There were also quite a few documents of importance that have simply been lost or "misplaced": the most obvious one is the document placing WWW in the public domain (of which I have a certified copy), and another one is my one-page proposal to make a study of networked hypertext, submitted to Gottfried Kellner, but of which I can't find my own copy! (R. Cailliau, personal communication, June 1, 2022).

Robert Cailliau's concern about these lost documents was also highlighted during the focus group: "I gave my entire […] five metres of documents to the CERN archives, and they were put in boxes, and they disappeared there actually. And I don't think I'm not sure this is something that someone should know ".

Therefore, we can say that, according to Robert Cailliau, two important documents have been lost: his proposal, which was written on yellow paper, and the original copy of the document from April 1993 that said the WWW was released for free in the public domain. This was something that Robert Cailliau brought up again and again when the focus group was talking.

In addition to missing documents, missing information can also be found within the CERN Document Server (CERN, n.a.–e). In the focus group, Cailliau said:

> There are hundreds of thousands of photos, some of which contain people, objects, and things that are important because they have become historically important. But of whom we have lost who is in this photo, when was it taken, where was it taken, what is the object?

Accordingly, under the term "disappeared", we not only include those records whose locations haven't been identified but also include those sources within the archive, and metadata, that need contextual information to be understood correctly.

9. Digitalisation and access

The final theme emerged during our research concerns how to access WWW collection, which is predominantly paper-based. Only two documents have been digitised and made accessible. On the one hand, the first Web proposal, Information Management: A Proposal, with notes by Mike Sendall (Berners-Lee, 1989). On the other hand, the Alexandria. Proposal for a European Centre for Networked Information Systems (Cailliau, 1993). As the access to sources is becoming an increasingly important issue in academia, our empirical research also wanted to find out whether CERN is considering strategies for digitising the archive to provide access even to researchers working in remote locations. Regarding this point, Jens Vigen during the focus group clarified that the critical point is funding:

> For the web collection, we don't yet have a sponsor to finance such a project, and it's a bit delicate to put everything online yet. *[…]* But I hope one day that we will eventually digitise those documents and make them freely available.

Therefore, to access the WWW collection's sources, researchers still need to travel to CERN in Geneva. Because of this, it is important to know how to have the permission to see the collection. In this sense, it is important to remember that Operational Circular N. 3 issued by the Personnel Division (1997) states that "The CERN Archives aren't public archives; they primarily serve as an information source for the organisation. However, access is given to people outside CERN upon request when there is "justified interest" (p. 5).

During the focus group, we then asked participants what is meant by the term "justified interest", and Jens Vigen explained: "[justified interest] *[…]* is somehow a code to express that this information will be of general interest to the public and

that it doesn't contain delicate personal information". In a few words, to access the archive, individuals must send a request to the CERN Heritage Committee. The request is taken into charge by chair Charlotte Warakaulle, Director of International Relations, who considers whether CERN has an interest in granting access.

Since, as a research group, we were granted free and complete access to the WWW collection and since we are digitising in Portable Document Format (PDF) with Optical Character Recognition (OCR) most of the records, we will try to negotiate with CERN if, how and when to place this material in open access.

10. Conclusion

This article provides a critical overview of the genealogy of the WWW collection at CERN dealing with its birth and development to understand what people, events, and decision-making processes converged on its creation and which people shaped and potentially still shape this archive. As mentioned in the theoretical framework, the building of an archive is a complex process, involving different actors, actants, and decisions which need to be remembered and reconsidered to help contemporary (Web) historians' understanding. This research investigates the spaces of flexibility where the collection evolved and shows that it is shaped by the efforts of multiple actors guided mainly by personal ethical values. Specifically, the WWW collection at CERN was formed through the collaboration of four people (Robert Cailliau, Dan Noyes, Ben Segal, and James Gillies) who, at different times (2001, 2013, 2021, and 2022), brought to CERN archivists the materials that form the WWW collection described in Figure 1. It must be remarked that according to our study the collection was created through the desire of a few individuals and not by a decision of CERN management. This is a characteristic that the WWW collection and the Web itself have, ironically, in common. Within those few people, Robert Cailliau plays an important role since 65 folders out of 99 contain his documents. Therefore, researchers should keep in mind that the WWW collection mainly reflects Robert Cailliau's views and experience, while missing other key players' perspectives. Even if Robert Cailliau played a major role, the overall action of preservation is collective and individual action is persistent in the shape of the collection, for personal reasons, for the ambition of writing a book, for the need of historical remembering.

In addition to the four people who packed the archive, this paper was also able to highlight the role of women, who remain invisible in Figure 1. These women have been paramount in the preservation of Web-related documents. It is worth mentioning the role of Robert Cailliau's wife, Susan Cailliau, who at the time of the Web's development was Division Record Officer in the division where the technology was being developed and who suggested to her husband to start preserving material since she realised how important the Web was about to become. Another key woman was Anita Hollier, the CERN archive manager who decided to create a space in the archives to preserve the material collected by Robert Cailliau and to compile and update the collection catalogue as new material arrived. This task has recently passed into the hands of Sandrine Reyes, another CERN archivist very active in preserving Web-related documents and who participated in our focus group. Finally, women have also been important in preserving

records outside of the official CERN archives. For example, Peggie Rimmer, Mike Sendall's wife, donated her husband's web archive to the British Library. New research is focusing on the role of women in the development of information technology. This article adds to those works by focusing on how important it is for women to keep Web archives safe, which makes them important figures in the history of the Web.

Our research also underlined the fact that CERN's WWW collection doesn't contain all the papers produced by CERN members during the Web era. Some of the missing documents are critical to tracing the history of the technology. For example, Tim Berners-Lee is the great outcast of this collection, and his papers, as well as those of others, can be found in the archives of private individuals or institutions outside CERN. Other documents, however, are still within the CERN archive but outside the WWW collection (such as Pier Giorgio Innocenti's collection). Consequently, it becomes evident that if researchers want to study the history of the Web by staying only in the WWW collection, they will miss many insights and perspectives that are outside of it. During the focus group, James Gillies explained that there are many stories surrounding the Web, and every person who has contributed to its development would like to tell their story.

This is just the first paper focusing on the WWW Collection at CERN and it mainly aims to describe and understand the reasons of its formation, in order to let Web historians and archival studies researcher have a clearer idea of the genealogy of this archive. Since the Web history is booming and since CERN is more and more willing to open its archives, new and more theoretical or source-driven studies are expected and welcome in the next future.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Martin Fomasi is a PhD candidate at the Institute of Media and Journalism (IMeG), Università della Svizzera Italiana (USI). His thesis project is part of the Swiss National Science Foundation (SNF) funded project "The origins and spread of the World Wide Web. Rediscovering the early years of the Web inside and outside the CERN archive." His project adopts a posture inspired by Science and Technologies Studies (STS) to retrace technical choices and influences during the early years of the Web.

Deborah Barcella is a PhD candidate at the Institute of Media and Journalism (IMeG),Università della Svizzera Italiana (USI). Her thesis project is part of the Swiss National Science Foundation (SNF) funded project "The origins and spread of the World Wide Web. Rediscovering the early years of the Web inside and outside the CERN archive" and is focused on how the Web was promoted and commercialised in its early years.

Eleonora Benecchi, PhD, is a lecturer and researcher (MER) at the Institute of Media and Journalim (IMeG), Università della Svizzera Italiana (USI). Her main research interests include Internet fandom, participatory culture with regard to the diffusion of pop culture contents through social media and the media consumption of young generations. She's the coordinator of the project "The origins and spread of the World Wide Web. Rediscovering the early years of the Web inside and outside the CERN archive."

Gabriele Balbi is Full Professor in Media Studies at the Institute of Media and Journalism (IMeG),Università della Svizzera italiana (USI). His research is focused on media studies with a historical and long-term perspective. . He launched and is responsible for the project "The origins and spread of the World Wide Web. Rediscovering the early years of the Web inside and outside the CERN archive". His last book is *The Digital Revolution: A Short History of an Ideology* (Oxford University Press, 2023).

References

Berners-Lee, T. (1989). *Information management: A proposal*. https://cds.CERN.ch/record/1405411/files/ARCH-www-4-010.pdf.

Bory, P., Benecchi, E., & Balbi, G. (2016). How the Web was told: Continuity and change in the founding fathers' narratives on the origins of the World Wide Web. *New Media & Society*, *18*(7), 1066–1087. 10.1177/1461444816643788

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press.

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide*. SAGE.

Cailliau, R. (1993). *Alexandria. Proposal for a European Centre for Networked Information Systems*. https://cds.CERN.ch/record/1378141/files/Alexandria%20proposal.pdf.

CERN. (n.a.–a). *The World Wide Web,* WWW *(Archives)*. https://cds.cern.ch/search?ln=it&cc=The+World+Wide+Web+%28Archives%29&p=&action_search=Cerca&op1=a&m1=a&p1=&f1=&c=The+World+Wide+Web+%28Archives%29&c=&sf=&so=d&rm=&rg=100&sc=1&of=hd.

CERN. (n.a.–b). World Wide Web [Draft]. https://sis.web.cern.ch/archives/CERN_archive/guide/IT/isawww.

CERN. (n.a.–c). Dan Noyes., https://home.CERN/authors/dan-noyes.

CERN. (n.a.–d). James Gillies. https://home.CERN/authors/james-gillies.

CERN. (n.a.–e). World Wide Web. https://cds.cern.ch/search?ln=it&as=1&m1=a&p1=&f1=&op1=a&m2=a&p2=&f2=&op2=a&m3=a&p3=&f3=&action_search=Cerca&c=World+Wide+Web&sf=&so=a&rm=&rg=10&sc=1&of=hb.

CERN. (n.a.–f). *CERN Computer Newsletter*. https://cnl.web.cern.ch/

Cook, T. (2011). 'We Are What We Keep; We Keep What We Are': Archival Appraisal Past, Present and Future. *Journal of the Society of Archivists*, *32*(2), 173–189. https://doi.org/10.1080/00379816.2011.619688

De Gregorio, C., & Manai, S. (2014). *Archives of David Owen Williams* https://sis.web.cern.ch/archives/CERN_archive/guide/IT/isadow.

Finlay, J. F. (2012). Five lenses for the reflexive interviewer. In *The Sage handbook of interview research: The complexity of the craft* (2nd ed., pp. 317–332). SAGE.

Fratila, A., & Sionis, C. (2006). Activating memories in interviews: An instance of collaborative discourse construction. *Discourse Studies*, *8*(3), 369–399. https://doi.org/10.1177/1461445606061880

Gillies, J., & Cailliau, R. (2000). *How the Web was born: The story of the World Wide Web*. Oxford University Press.

Hollier, A. (2005). Arrangement and Description of the CERN Archive. *Business Archives*, *89* https://businessarchivesjournals.org.uk/Filename.ashx?tableName=ta_businessarchives&columnName=filename&recordId=65.

Hollier, A. (2007). *Les archives dans un grand organisme de recherche européen*. https://cds.cern.ch/record/1099782/files/cer-002753625.pdf.

Hollier, A. (2011). Introducing: The CERN Archive. *Viewpoint (Newsletter of the British Society for the History of Science,) 95*(1), 1.

INRIA. (n.d.). *The history and future of the Web*. https://www.inria.fr/en/entretiens-croises-pour-les-30-ans-du-web.

Jardine, B., & Drage, M. (2018). The total archive: Data, subjectivity, universality. *History of the Human Sciences*, *31*(5), 3–22. https://doi.org/10.1177/0952695118820806

Latour, B. (1994). On technical mediation – Philosophy, sociology, genealogy. *Common Knowledge*, *3*(2), 29–64.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Harvard University Press.
MIT. (n.d.–a). Tim Berners-Lee. https://archivesspace.mit.edu/repositories/2/archival_objects/205201.
MIT. (n.d.–b). *Tim Berners Lee – World Wide Web –MLD*. https://archivesspace.mit.edu/repositories/2/archival_objects/205472.
MIT. (n.d.–c). Berners-Lee, *Tim. World Wide Web Consortium, 1994–1994* https://archivesspace.mit.edu/repositories/2/archival_objects/94582.
MIT. (n.d.–d). Berners-*Lee*, Tim, 2007–2011 https://archivesspace.mit.edu/search?utf8=%E2%9C%93&op%5B%5D=&q%5B%5D=Tim+Berners-Lee&limit=&field%5B%5D=&from_year%5B%5D=&to_year%5B%5D=&commit=Search.
Morgan, D. (1998). *The Focus Group Guidebook*. SAGE Publications, Inc. https://doi.org/10.4135/9781483328164
Noyes, D. (2013). *1999 backup of TBL's NeXT hard drive surfaces*. https://first-website.web.CERN.ch/first-website/node/27.html.
Personnel Division. (1997). *Operational Circular* N. 3. http://cds.cern.ch/record/1202773/files/CERN_Circ_Op_en_03.pdf?
Potter, J., & Hepburn, A. (2012). Eight challenges for interview researchers. In *The SAGE Handbook of Interview Research: The Complexity of the Craft*. SAGE Publications, Inc. https://doi.org/10.4135/9781452218403
Puchta, C., & Potter, J. (2004). *Focus Group Practice*. SAGE Publications Ltd. https://doi.org/10.4135/9781849209168
Roulston, K. (2016). Issues involved in methodological analyses of research interviews. *Qualitative Research Journal*, *16*(1), 68–79. https://doi.org/10.1108/QRJ-02-2015-0015
Speer, S. A., & Stokoe, E. (2014). Ethics in action: Consent-gaining interactions and implications for research practice. *The British Journal of Social Psychology*, *53*(1), 54–73. https://doi.org/10.1111/bjso.12009
Vaughn, S., Schumm, J. S., & Sinagub, J. M. (1996). *Focus Group Interviews in Education and Psychology*. SAGE.
Vigen, J. (2022, June 2). *Focus Group to write a paper on the origins and development of the WWW collection*. [Personal Communication]
Wertz, F. J. (2011). (A c. Di) *Five ways of doing qualitative analysis: Phenomenological psychology, grounded theory, discourse analysis, narrative research, and intuitive inquiry*. Guilford Press.
Wilkinson, S. (1998). Focus group methodology: A review. *International Journal of Social Research Methodology*, *1*(3), 181–203. https://doi.org/10.1080/13645579.1998.10846874