TECHNOCULTURE

Digital forensics: A detective in the archive

Episode 5

Full transcript

Guest: Thorsten Ries [Thorsten]
Host: Federica Bressan [Federica]

[Federica]: Welcome to a new episode of Technoculture. I'm your host, Federica Bressan, and today my guest is Thorsten Reis, who is a Marie Curie fellow at Sussex University in the UK and currently associate researcher at Ghent University in Belgium at the Institute of Modern German Literature. Welcome, Thorsten.

[Thorsten]: Hey, thanks for having me, Federica. It's exciting to be here.

[Federica]: You're currently in the UK. What kind of research do you do there, and tell us a little bit about the keywords of your work.

[Thorsten]: Yes, my current research project here in Brighton at the University of Sussex — well, at the School of History, Art History and Philosophy and the Sussex Humanities Lab — is about digital forensics in the historical humanities, and a very short version of what, well, would summarize my research is, I'm interested in the digital materiality of the digital historical record, so what is hidden in the digital materiality of hard drives that become part of archives that are being committed to the archives, to the libraries, what happens to our digital heritage that we leave behind maybe on social media, maybe committing them to archives, and from multiple sources — literary sources, but also historical sources and historical personal archives.

[Federica]: We will explore this concept in the next hour, but there's this adjective that really caught my attention from the very first time that I saw your research page. Now, in audio forensics, that does mean that you kind of have to prove in court if a recording comes from this environment or this voice is or not the voice of that type of person, but I understand in your research the adjective ‘forensics’ does not imply actually any work in court, so what does it stand for?

[Thorsten]: Yes. My research is not about criminal activity. It is about historical preservation, and it is about historical appraisal of primary sources, and the word ‘digital’ or the term ‘digital forensics’ comes into play because essentially what we are using for these goals is methodology and tools that actually have been developed in the digital forensic realm. It was actually law enforcement that gave us the tools that we use today to analyze, for example, hard drives for philological reasons. So I'm giving an example. In the past, I've been working a lot with the personal digital archive of a German author whose name is Thomas Kling. I have been analyzing his hard drives after I preserved them with what we call forensic imaging, which is nothing else but a bit-precise copy for lost and hidden versions of his writing. So there was basically a very philological endeavour, finding draft versions of a poet which would have been otherwise lost if we did not look at the original hard drive, if we did not look under the surface of the files as they would appear in a normal word processor.

[Federica]: In this case, what tool did you use that came from law enforcement?

[Thorsten]: Multiple. It's a whole... Actually, it's a whole range of tools that you use, and one of the easiest-to-explain tools probably is our data recovery tools which basically recover data that has been deleted, either in fragments or as whole files, if they have not been overwritten properly. So that is one of the tool categories that are very important in this kind of research.

[Federica]: So files that were deleted intentionally by the person who was using the memory storage device, and you kind of go and look... It's like, you know, looking into the trash bin of a person who was murdered or simply left, something like that.

[Thorsten]: Well, no one has been murdered in this case.

[Federica]: [laughs] I'm sorry, I just, the word ‘forensics’ just calls all these images in my mind that you're like a detective, you know, this philologist detective. It's just such a fascinating thing.

[Thorsten]: Well, philology has always been a bit of detective work, so the semblance makes sense or the comparison makes sense, but one important, one very important prerequisite, of course, of this kind of research is that you have permission to do this, so this kind of access to the digital archive requires, of course, explicit permission and explicit consent of a data subject for the investigator, for the researcher, to look into these archives, into these digital archives, in this way. So there have to be good agreements and good contracts and be, made beforehand before you actually go and do that.

[Federica]: Can you tell me a successful story of sort of detective work that you've done, something. . . You know, I assume that sometimes you don't find much. Have you ever found something that then was really interesting to study?

[Thorsten]: Yes. Mostly, I do. There is. . . Well, you're presupposing that often or mostly you don't find much, but actually the experience is that on our digital writing tools, a lot of traces of what we did with them in the past actually remain there and can be, may be found later. This situation of what can be found and which materiality these traces have change historically with all technical versions of the operating systems and the writing tools that we use. And you asked for an example. I have been studying Thomas Kling's hard drives for a long time, and some of the most dramatic findings that I had, or the findings that moved me most, were from the period when he already, when he was working on his last poetry and essay volume and he knew that he was going to die very soon. And during that time, his computer crashed very badly, and we were able to actually recover large portions of this hard drive with some laboratory effort, and it was actually possible to recover quite a lot of his work that he did until the point that this computer crashed and he had to start over with some texts and where he sort of had to start over at an earlier point from which he had backups. So it was really traveling back in time into a very dramatic computer crash where someone has been, where this author has been really writing against time.

[Federica]: But he recovered portions of those data when he was still alive or afterwards?

[Thorsten]: Afterwards. This drive has been committed to his archive.

[Federica]: So were you trained as a digital philologist? How did you embrace the digital in your philology?

[Thorsten]: I am professionally trained as a philologist. I studied German literature in Hamburg and a little bit in Baltimore and then I came later to, again, to the university to finish my PhD project. And while I was working on my PhD on a 20th-century German author (specifically, about his writing process and about his notebooks and about his sketches and his drafts), I was constantly already wondering: 'Okay, I'm doing this work now. This is really interesting, and I think this is still interesting, but this work, like this poet's work, nowadays, is being obviously done on computer.' It's like, what remain — and the question was, what remains of it? Like what are the traces? And that was the point when I started being interested in digital forensics, essentially. I started. . . I always have the suspicion that our writing, that our digital writing tools are basically [unclear 00:10:57] on us and save a lot of stuff that we didn't want to save and that it's not all gone just because we delete something or because we switch off our computers. And I had the feeling I want to look at it and do something good with it for philology, and it happened that I was just working on a talk about this when I discovered Matthew Kirschenbaum's book, Mechanisms: New Media and the Forensic Imagination, which has been just published two years before and was not very known in Europe at that point. And it was actually not so hard to find this book at that point.

[Federica]: What year was that?

[Thorsten]: Kirschenbaum's Mechanisms book was published 2008, and my talk that I gave in Frankfurt was in 2010.

[Federica]: So since then, how much did you have to learn technically? You know, to run, I assume, some algorithms that [powers 00:12:12] the hard disc drives and retrieve this information, how much literacy in computer science do you need to have to do your work?

[Thorsten]: It's a constant learning process, essentially, and it's less and... Well, the thing that I always say to encourage other researchers to do this kind of work — because there's so much work to do; I cannot do that all by myself — I always try to encourage people and say, 'Hey, you need some you need some skills, but you don't really have to become a coder for this.' You have to learn a lot about historical computing and how these traces got onto that computer, and this mostly [unclear 00:13:02] code something, but you have to understand how a historical application worked or how a historical operating system worked, with which granularities it worked, with which metadata it worked, which bugs these specific systems had, which weird ways of handling their memories these historical systems had. This is something that you learn while you go, and this is something... And it's a constant learning process. Just today, I was sitting down and had to learn about how WordStar has been handling its memory.

[Federica]: That sounds a lot of technical knowledge to me. You say you learn as you go, but that belongs to media archaeology, basically. Is this a field that's so developed, established already so that a person like you or one of your colleagues would specialize in a period, like the '80s or the late '70s, or you kind of have to know it all a little bit?

[Thorsten]: I presume that at some point people will have to specialize on certain periods, I guess, because it become — because while we progress in time and in history, the history of computing and its digital material traces, its forensic traces, will become vaster and vaster, and the history becomes in many ways more complicated in the present by, for example, encryption, mobile devices, and so on and so forth, which is something that my present project is also about.

[Federica]: Right. Okay, so not just computers but also the smartphones and tablets of the people whose work you're studying?

[Thorsten]: Potentially. Well, the subjects that I am looking at right now, that is not really in itself about mobile computing. These sources have been produced with plain computers, laptops, as we know them. I'm looking at the personal digital archives of Hanif Kureishi, the author, the London-based author, of a London-based tech journalist, Glyn Moody, who was kind enough to give us access to his personal archive as well, and into the born-digital parts of The Mass Observation Archive, which is being preserved constantly at the University of Sussex. And to the question about mobile devices, I am looking at that in the training part of my project. I'm going to learn how to extract data and memory from storage media that is usually built into mobile devices and very modern laptops, for example, which are solid-state drives or flash memory, which is a totally new game for digital preservation. Did you ever ask yourself how to preserve the memory of a historically important cell phone for the archive? This came up in the Library of Congress a few years ago when a journalist who was embedded, I think, in Syria gave his cell phone with which he was taking photos from the battlefields, as I remember. He gave his phone to the archive, to the Library of Congress, and they had to preserve it, and it was quite difficult to preserve a cell phone that was not switching on again. So how do you get the memory out there? How do you preserve this? Do you just preserve the object, or do you think you can get down to the memory and preserve that? And that is something that I'm very interested in right now.

[Federica]: Do you work alone, or is it always teamwork to carry out such task of recovery?

[Thorsten]: That depends. That depends on the task and on the situation. There's a lot of things that I do just by myself and where I do the analysis and where I do the preservation just by myself, but there are also cases where I also need help and where I have to work together with others. For example, when a hard drive is very badly damaged, I'm also just saying, 'Okay, this is a case for a forensic lab. They have very different machinery there. You need [different 00:18:55] equipment for that.' And so then I also work together with others.

[Federica]: There is something I want to ask and that is, is your work preliminary to the philologist's actual work of analyzing the sources? Once you have retrieved the files, the documents, the data, then you do something with them, and technically the recovery process is just preliminary to that, or is this included in the definition of a philologist's work?

[Thorsten]: That depends a little bit on how you see philology and also what you think archival work is. I think that born-digital sources are shifting our whole framework of how we see the archival task and the historian's and philologist's task. Right? Today, when a digital archive is being committed to a memory institution, archivists cannot just take it and put it on the shelf. They have to process it first, and there, some of the work that I am doing as a philologist has to be done already. There's some recovery potentially going on or already thinking about the potential of recovery, meaning: What, maybe, do we have to effectively delete before we can really commit this to the archive and before we preserve this? And so these are questions that require actions by archivists that are also part of the philologist's work as well. On the other hand, my own work is spanning from preservation to analyzing the primary sources up to thinking about how we could include these kinds of materials and scholarly editions, for example. How do we present these to a scholarly public? And this is... So my work, I think, is spanning over and at the same time at the interface of archival work and the work of philology and scholarly editing.

[Federica]: Let me touch on the topic of authenticity for a moment. Philology is concerned with determining whether a document is authentic or not. Reliability and a bunch of other issues follow from that. I wonder how that translates to the digital domain. Is digital philology, first of all, concerned with authenticity, and if so, how can you determine whether a digital document is authentic or not?

[Thorsten]: That is a very... It's a very interesting and very broad question. Of course, well, while we are all confronted with constant news production about fake news, about, well, about digital forgery, about news even that can be fabricated where even footage can be fabricated, videos can be totally, well, as they call it, doctored, well, you might have seen videos where Barack Obama has been like made say something else than he actually said during a video, and they just they just changed his facial expressions so he actually said those other things just made from another video. So in these times, of course, the question of authenticity of digital sources, of course, comes up quite quickly and in almost every conversation, and the most, or the core of the answer is custodianship and curatorship. The most important factor in authenticating digital sources is, someone has to determine at a certain point of time in a certain place in the world that this source actually exists and that is its provenance. That has been checked. And from then on, guarantees that this source, this digital file, is authentic in this very specific bitstream form, meaning these bits in this sequence actually represent the same source that has been secured and authenticated at this and that point of time and in this and that archive — which guarantees the authenticity. So there are means to do so.

[Federica]: Is the authenticity of most documents you deal with kind of known from the very start because they come from the hard disc drive of the person that gave it to you?

[Thorsten]: To say it in a technical way which might sound a little bit like law enforcement again, there is a so-called chain of evidence which can be established from a physical storage medium which has a specific serial number, which is a physical object that can be located in a, which is located in a very specific place, in an archive, and from which the data has been derived in a controlled way and it can be proven that this data is identical with the data on this storage media at this and that time. So yes, there is a chain of custody, as we say, between the historical physical source and the bitstream-identical data.

[Federica]: We focus on digital media a lot. We've mentioned digital archaeology. We work in the digital humanities. But is it accurate to say that you don't work with archives that are all digital? You have analog documents too, so actually archives are hybrid. They have both components, analog and digital, and as a digital scholar, how do you deal with this complexity?

[Thorsten]: My experience is that we have seen a transition to more and more digitally literal authors that handle their digital writing tools in a more and more routine way up to digital writers who actually code their own literature. But at the same time, I have so far not seen any purely digital archive. They are pretty much all hybrids, and that also makes it especially interesting. I mean, I am currently finishing a book where two large chapters are actually about hybrid writing processes where I am trying to analyze examples where the authors were switching between analog writing tools such as notebooks or revisiting their prints, their printouts, their hard copies by hand and re-enter into the digital writing process, taking their notes into their digital documents or putting their revisions back from paper on into the digital tool, into the digital environment. So these switches between analog and digital writing is something that I'm really, really, interested in. So this happens a lot, and most archives are hybrids.

[Federica]: How large is the community that works with digital forensics? Is it a fairly large established community, or the whole field is still seen like a niche, like a subset of main philology?

[Thorsten]: This is a very, this is an interesting question, and we have the weird situation at this moment that in the archival sector, we have a lot of specialists for legacy, historical, digital primary sources and hard and software for the preservation side who are very proficient and also analyzing the stuff and who educate themselves constantly how to actually do the philologist's work, but they don't do it (they are archivists), but they would know how to do it. So there's a lot of knowledge being accumulated on that side. But on the other hand, philology and history. . . Jane Winters recently has pointed that out on the DH Benelux Conference. History studies and philology is really slow in developing a field of research around it, and I think I can say that Matthew Kirschenbaum and Doug Reside and some others in the U.S., and Luciana Duranti and Corinne Rogers in Canada, are important in this field on the scholarly side. Here in Europe, I think I can say that I'm still a bit a pioneer.

[Federica]: Can you just translate the methods of, let me say, traditional philology to this type of recovering hidden, lost, deleted sources digitally, or you need to find, you know, new methods that are specific to this type of material?

[Thorsten]: Well, within philology, the method is, of course, new because it's digital and because it's working with digital primary sources. The methodological sets are, I think, compatible with each other because philology always had to do a lot with forensic practice, and forensic practice had a lot to do with the scholarly practice of securing bibliographical evidence, securing material evidence. Archeology, philology, bibliography are fields of research that have always had a certain relation with forensic practice in court. Therefore, it is actually quite easy to translate, for example, mechanisms and methods to cite a piece of digital evidence in court to a method to cite a digital draft in a research publication. This is pretty much, you can just do it the same way, essentially, and then it's a sound method. So there is... You can certainly translate methodology from one field to another here. What is a little bit more difficult, maybe, is to translate parts of digital forensics from the past into a philological practice of the present. What does that mean? Digital forensics normally has the interest to solve cases that are very recent, mostly, so they mostly deal with very current technology. They don't care so much to solve cases with machines that are 30 years old. Right? So the preservation even of tools and methods that are able to analyze legacy hardware and software, that is actually a task that archival science, philology, bibliography will have to deal with in the future because forensic science or digital forensics doesn't do it [for them 00:33:58], but that's also a chance. Right? I mean, this is really a working field for the humanities.

[Federica]: I'm interested, there must be an ethical aspect to this. Is it as easy as what you shortly explained earlier — that is, I will look into somebody's hard disc drive if they let me do it, and I won't if I don't have permission to do it? How is the ethical, yeah, the ethical implications of your work? Is it complex or easy?

[Thorsten]: Both at the same time.

[Federica]: I like that you never have an easy answer. Like, 'It depends.' You know. Okay. Okay. Tell me about it.

[Thorsten]: I can explain. I can explain. In principle, it is easy. As a philologist, I am asking before I do something. I'm asking the data subject before I'm doing something, and I'm asking for so called for so-called 'informed consent', which means I describe in large detail what I'm going to do, what can be the outcome, what could I possibly see, to the heirs or to the data subjects themselves, and only then I start doing something. And you can make very good... I think what my experience is, you can make very good agreements that build trust and where you can limit the impact on the data subject in terms of privacy protection, for example, to a very comfortable minimum so that the data subject can be comfortable that their rights are being respected, that their privacy is being respected and so on and so forth. You can limit your investigation to the relevant portions of data and so on and so forth, but of course, these investigations have multiple levels, and there can... Many things can come up. Some document can come up that you did not want to see. That might be a very personal letter, or that might be a letter or an email that contains privileged communication that you did not want to see, so there is definitely an ethical component there, and this is also something that you have to negotiate beforehand. What is your intention? What would you do when you see that? Mostly it is that you inform the data subject about, that you saw this, that you're keeping discretion on it, and so on and so forth. Well, and there's multiple levels involved. A colleague of mine, James Baker, recently cited cases of personal digital archives where apparently audio files were included that probably have been downloaded via Napster, where, you know, at that time, it was gray zone and nobody really could say whether it was legal to have these audio files or not, to stream them or not stream them, have them on your hard drive. That was like a gray zone. And these are obviously ethical issues that you can encounter, but this is the same with all archives. I mean, it is... Also, paper archives contain material that the data subjects or the archival subjects have not deliberately committed to the archive but that just slipped in with the folder that they brought there.

[Federica]: That's true, although I think that hard disc drives of personal computers or smartphones take that to the extreme because we use them for leisure, for work, so I get very nervous when we talk about looking into somebody else's hard disc drive. It's like checking my Google search history, you know, my watched YouTube videos. I get very nervous about it, so I insist on this a little bit. I have to ask again. I'm thinking the case of the poet. You're interested in previous versions of how he was writing the poems, but then if you look into his hard disc drives, you may find, yes, letters to family members or lovers or debts he had, you know, all that personal, sensitive material. So as a philologist, I really don't know the answer to this. I'm really asking, do you stay on track (like only focus on the poetry, so to speak) or no, actually it's more holistic even the philological approach is like, 'Everything informs me about who the person was and therefore also sheds light on his work'?

[Thorsten]: Well, this question is fairly specific, and so is my approach. I am working often even with living data subjects. In the case of Thomas Kling, Thomas Kling has already passed away, so I was negotiating with his heir about these things. My approach is that I methodologically limit the scope of my search and the... I am in a relatively comfortable position, as a philologist who is interested in writing processes because keyword search is a very good tool to limit the scope of what you see, so if you use good keywords or good key phrases, you really get only versions of a text, right? Or at maximum maybe an email where the author cited himself and said like, 'Recently, I wrote this and that, and I think this applies to your situation as well.' You know, something like that. So technically, I am able to limit my search and not just, you know, wildly search around. I mean, and this is part of the deal. This is something that I always make part of the agreements that I have with authors, data subjects, and heirs: that I keep the scope of my search that specific.

[Federica]: Hmm. I think that I have a weird mind and I'm putting all these words together, the forensic, the detective work, and looking into somebody's hard disc drive, so I'm probably having just a set of scenarios in my mind that just do not apply. They're completely fantastic, as opposed to the real work you do, but I'm sorry, so... There must be some of that, anyhow. Let me say this. So far, you've been mostly speaking of cases where the data subject is alive, so you negotiate with them. I understand that legally, when the person has deceased then it's the heirs. That still makes me a bit uncomfortable because it's something extra personal. So, I mean, even my own family members, I... That's weird. Well, I was going to say, I was just going to say that probably I wish that my own hard disc drive is just destroyed completely when I'm gone because even my family members don't know what's in it, so how can they give permission to look into it? At least I know what I have done and I can say, 'Yeah, go ahead.'

[Thorsten]: I understand the uncomfortable feeling in your case, but you are... Well, you're constructing a situation where you have no control over the archive that is being committed to a memory institution, where you cannot know that this would happen, like you are constructing a situation in which your heirs alone have to decide about this. Let me tell you just how the normal case that I encounter normally goes. Normally, it is the author's themselves who decide that their digital archives will be part of their archive and that they have to be in their archive and then it is their heirs who then, well, physically put it there and integrate it into the archive and do the administration, the management of the archives, and take the decisions on behalf of the data subject. So there is an important component in there where the creator has already decided that this is something that has to be included into the archive.

[Federica]: Mm-hmm. You've said that oftentimes authors will volunteer their archives for you to study. I was wondering, can also regular people share their own material, whatever it is, for the good of science, so to speak, to increase the body of documents available to refine the tools of science?

[Thorsten]: Yes, this actually happens, and it's actually... One of these archives is actually part of my present project here at University of Sussex. It's the Mass Observation Archive or the Mass Observation Project Archive, which is a very long-running project of... Well, I should say it is a grassroots history project where so-called observers all over the UK who volunteer to be observers commit records about their observations of the year on multiple levels to the Mass Observation Archive and thereby, well, because there's so many, create a grassroots history of the UK for many decades already.

[Federica]: I'm not sure I understand. What is shared?

[Thorsten]: The observers share their observations, their general observations. They might be personal. They might be political. They might be social. They might... They can have whatever scope. They commit these observations in form of a letter, a record, to the Mass Observation Project Archive on a yearly basis.

[Federica]: So I'm a citizen of the UK and I just write a letter to say how I feel about Brexit.

[Thorsten]: I guess there's a lot of Brexit reporting in the [Mass Observation Project Archive 00:47:20] these years, yes.

[Federica]: Who runs, who hosts, who supports this archive?

[Thorsten]: The Mass Observation Project Archive is currently hosted at the University of Sussex, or more specifically at the archive called The Keep, which is associated with the University of Sussex. And what is really interesting about the Mass Observation Project is that it actually goes back to its foundation in 1937, so that's pretty old. And there is an interruption in their work. Their work ended in the mid-'60s but then was revived in 1981, so this is really a long period of reporting on the grassroots level on everyday life, on everyday opinions of people in the UK about important historical and social developments.

[Federica]: So the first part of this archive must include a lot of analog documents. Do you know anything about the process of digitization to keep it up to date, to feed into the, you know, processable datasets today?

[Thorsten]: Yeah, this is exactly what I'm interested in. Obviously, a large part of this archive has been created analog, and I think they also work on digitization of these analog sources, but what I am more interested in are the more digital sources, like I am interested in the period in which the creators or the observers started submitting their reports in digital form and in different digital formats. What do those formats, those digital formats, tell us? Are there interesting artifacts in their reports? Do those reports reveal in their materiality another history of the digitization, of the process of digitization, of the UK, for example, by, you know, very simply the percentage of digital reports that have been submitted at certain times compared to the analog ones or how many, in which way did the observers on average use formatting in their texts, which becomes more prevalent in the course of time when it becomes easier, or in which formats created by which text processors did they submit their reports? And so on and so forth. I think, or what I'm interested in is, how does this reflect the digitization of UK society over time?

[Federica]: Speaking of digitization of society, which is a theme clearly very dear to this podcast, can you tell us a little bit about how it all began? You know, how far are we? I know it's a complex question and the situation is not even across the world, but, you know, can you tell us something about, hmm, at least in your experience, what were the early signs of digitization, so to speak?

[Thorsten]: Well, you do give me a chance to first start out with the area or the era that that I have been studying a little bit more, and apart from the work that I'm doing right now with the Mass Observation Project, I have been mostly looking at personal digital archives that come from end of the '80s and the '90s up to the millennium years, basically. That's the time period that I looked at, and they partially span into the present-day period. So when I work with living data subjects, that, of course, borders upon our current time. And if you ask me how far are we with digitization, I think this is a very broad question, and I don't know whether I have a better answer than your own feeling is, our society is pretty digitized, and it became an indispensable medium of everyday life — political culture, literary culture.

[Federica]: Thank you so much for this answer and all your answers, for sharing your knowledge. Unfortunately, our time is up, so I want to thank you for being with us and wish you all the best of luck with your project.

[Thorsten]: Thank you.

[Federica]: Thank you for listening to Technoculture! Check out more episodes at Technoculture-podcast.com or visit our Facebook page at Technoculturepodcast, and our Twitter account, hashtag Technoculturepodcast.