III Digital Strategies: Opportunities and Synergies

Ulrich Fischer

Digital Reconstruction of the Cologne Fragments – How, with the Help of AI, 'Cologne Flakes' Are Turned Back into Complete Items

Abstract: After the devastating collapse of the Cologne City Archives in 2009 approximately 95% of the archival holdings were salvaged – with much of the material largely intact but some of it highly fragmented. During the process of reviewing these items ('salvage capture'), which will soon be concluded, millions of 'puzzle pieces' have been recovered from all types of holdings. As with all salvaged archival material these fragments are intended to be restored so as to render them usable again in the medium to long term. However, manually reassembling these fragments which are soiled, completely jumbled and mechanically damaged is out of the question. Furthermore, the available staff need to focus their conservational and archival resources on working with the archival material that survived intact. For these reasons a separate workflow was identified for processing the fragmented archival holdings. At the recommendation of the Fraunhofer Institute for Production Systems and Design Technology (IPK), the Historical Archive of the City of Cologne cooperated with its partner MusterFabrik Berlin to develop a technological solution and carry out the appropriate procedures to reassemble these so-called 'Cologne flakes', virtually at first. Since 2020, the 'Digital Reconstruction of Fragments from Cologne' project has been under way on a massive scale. In combination with the employees' specialist skills and knowledge of the holdings the Historical Archive of the City of Cologne has found the deployment of artificial intelligence to be an important tool for reconstructing even these severely damaged archival materials.

Article note: This contribution was produced together with Bettina Schmidt-Czaia, Wolfgang Meyer and Hanna Kannengießer. It is a slightly revised and updated version of the text of the oral presentation, reflecting the status of the reconstruction work on the Cologne City Archive at the end of 2022.

Ulrich Fischer, Historical Archive of the City of Cologne, Cologne

https://doi.org/10.1515/9783111386713-008

It is undisputed that the collapse of the Cologne City Archive in March 2009 was a decisive date for the handling of written cultural heritage.[1] The impressions created by the destroyed building and also of the joint salvage and immediate rescue efforts acted as a catalyst to accelerate the already existing tendencies towards a comprehensive understanding of preservation and conservation.[2] The focus in conservation shifted from individual items to entire holdings and collections; concepts for the cleaning and packaging of large quantities of archival material, overarching digitisation efforts and cooperation between institutions in preservation and emergency prevention became the focus of professional discussions – and consequently are among the funding priorities of the Coordination Office for the Preservation of Written Cultural Heritage (KEK).[3]

Of course, all these aspects play a role in the reconstruction of the Historical Archive of the City of Cologne.[4] The innovations in the field of reconstruction that have shaped the work include the packaging strategy based on the newly-developed Rhenish archive box [‘Rheinischer Karton’][5] quantity processes for dry and wet cleaning, the extensive employment of assistants in restoration and the registration of salvaged items, barcode-based logistics including dynamic instead of inventory-based storage and, finally, the digitisation of processes that allow the work of archivists and conservators to be controlled and documented seamlessly in an additional module to the archive database.

Much of what was developed and introduced for the reconstruction work at the Cologne City Archive in the first decade after its collapse is currently being adapted to the conditions in our new building at Eifelwall. Here the focus is on the

1 In the Coordination Office for the Preservation of Written Cultural Heritage’s (KEK) first publication, for example, the collapse of the Cologne City Archive is cited on the first three pages (in the introductory words of the then Federal Government Commissioner for Culture and the Media (BKM), Bernd Neumann, the then Chairwoman of the Cultural Foundation of the German Federal States, Isabel Pfeiffer-Poensgen and in the body text itself) as the driving force for the creation of this institution and the search for national strategies for the preservation of written cultural heritage, see <https://www.kek-spk.de/publikationen#anf%C3%A4nge-der-kek> (last accessed 18 October 2023).

2 Thus as early as 2009 in the position paper *Zukunft bewahren* by the Alliance for the Preservation of the Written Cultural Heritage, see <https://www.kek-spk.de/publikationen#mehrsprachige-ver%C3%B6ffentlichungen>, 7 f. regarding the archives (last accessed 18 October 2023).

3 See <https://www.kek-spk.de/foerderung> for various examples of the aforementioned funding priorities (last accessed 18 October 2023).

4 See Fischer and Späinghaus 2019 for further literature.

5 The Rhenish archive box is a slip-lid cardboard box of folio format and a height of 12 cm, which serves as the basis for a packaging system to which the shelf dimensions in the new building at Eifelwall have been customised.

consolidation of procedures as part of a general process management as the size of the task and the diversity of the individual measures required for success necessitate intensive coordination, the definition of processes, knowledge management and above all coordination at the interfaces of the organisational units involved. All procedures that are adapted or newly developed must be orientated towards the main objectives. In this context the emphasis is on making archive material available for use.

For the greater part of the approximately 95% of our archive material recovered between 2009 and 2011 we have procedures in place to achieve usability – digital or analogue, with more or less lead time. This is the case for all those archival objects that were salvaged (largely) intact and which could be made retrievable by reconnecting them to an entry in the digital finding aids during one of the various phases of registration. In addition, salvaged archival units that could not be linked to the digital finding aids can be retrieved as they were described in the registration process.[6]

But what happens to the rest, to the fragments? Here we distinguish between the 'formation fragments' and the 'puzzle fragments'. The former are individual pages or salvaged archival units with several pages but which no longer form a complete archival item and whose allocation to cataloguing units remains a major challenge. Much more problematic still, however, were and are the so-called 'puzzle fragments', i. e. parts of pages that our careful salvage and sorting teams were able to recover from the rubble. For these, the first responders coined the somewhat offhand term 'Cologne flakes', and they collected them by the boxful (Figure 1).[7]

In addition to smaller quantities of fragments in normal recovery boxes, entire groups of fragments were collected in specially created boxes when all the recovered items were recorded. In total, several million puzzle fragments were salvaged after the collapse.[8]

This amount of fragmentation of our archive material is not surprising; in many cases, it was the enormous shearing forces that destroyed entire files in the course of the collapse. Additional fragmentation resulted from the exposure of archival materials to water, in some cases for months, and also from security and evidence preservation measures on-site, in particular drilling and other measures to explore the subsoil at the site of the collapse. Considering the extent of the destruction it

6 More than half of the estimated 1.5 million salvaged items can be found and ordered for use, allowing for an appropriate lead time for their conservation and archival processing.

7 This term first appears in a report by Dietmar Bartz for the *taz*, see Bartz 2009.

8 It is still not possible to determine the exact number, as the fragments were collected but not counted in the processes to date.

Fig. 1: Fragments from the Historical Archive of the City of Cologne. Photo: City of Cologne

is actually astonishing how many archival materials of all types survived the collapse intact. Essentially, the fragmentation affects all holdings that were stored in the collapsed storage rooms. The respective extent results from various externally specified variables. The storage location of the holdings or parts of the holdings in the collapsed storerooms play a role, as do the type and quality of the archival packaging, the materiality of the objects and the salvage location from where the objects were salvaged.

Assembling the fragments (puzzle fragments as well as formation fragments) is a special task in the context of reconstruction. They are the weakest link in the reconstruction process so to speak, as the restoration and archival challenges are particularly concentrated in the small, torn objects. In other words, reassembling an indexed archival item that has been broken into puzzle fragments to make it usable again in its original form demands many times the effort required for physically intact salvaged indexed archival items. Therefore, the question is: why does the Historical Archive deal with these particularly problematic cases – and how can the fragmented items best be reprocessed?

The question of why is clearly answered by the North Rhine-Westphalia Archive Act: paragraph 2 section 7 stipulates that ‘archive material […] shall be preserved

[and] maintained [...]'.[9] There is no doubt that the fragments created by the collapse are archival material as they had previously been checked for their archival value by an archivist and subsequently transferred to the holdings of the Cologne City Archive;[10] Hartmut Weber, the former president of the Federal Archives, includes the reassembly of the fragments in his expert report for the Cologne Regional Court and reckons that there are easily three million of them. However, not all of the 'Cologne flakes' are fragmented archive material. The puzzle fragments include fragments from the library and material from the archivists' offices and neighbouring buildings. In many cases, however, a distinction between these categories cannot be made on a puzzle fragment alone and thus, in case of doubt, the puzzle fragments found are further processed together.

While the overwhelming majority of the archive material affected by the collapse is owned by the city it is well known that some of the bequests and collections are permanent loans. The Archives Act and its relevant regulations on the preservation and restoration of documents worthy of archiving also apply directly to these holdings; this means that even if the fragments from these holdings could be reliably distinguished from other archive materials the obligation to process them further applies analogously to them as to records of municipal provenance.

In addition, not only the City of Cologne but also the owners of such permanent loans have claims against those responsible for the collapse of the archive. This means that mitigation and preservation measures must be applied equally to all the material, whether municipal property or permanent loans, whether fragmented or intact. This is understandable as the decision in favour of the inclusion of succession property in the archive – be it as a permanent loan or with a later transfer of ownership to the City of Cologne – was and is made for the same reasons as for municipal administrative records; the aim is to secure unique, archive-worthy heritage for posterity.

In view of limited personnel resources, however, it is an important goal to process the 'Cologne flakes' manually as seldom as possible as the available capacities can be used more efficiently for restoring unfragmented archive material. However, if the preservation, identification and assembly of fragments from the collapse are essential from a legal point of view but at the same time the personnel, logistics and space required for this 'conventional' processing are not available, the only remaining option is the development and introduction of technical procedures for the processing of the fragments. The Cologne City Archive has followed these

9 <https://recht.nrw.de/lmi/owa/br_bes_detail?sg=0&menu=0&bes_id=13924&anw_nr=2&aufgehoben=N&det_id=320558> (last accessed 18 October 2023).

10 For the following, see Fischer 2015.

guidelines in the development of its 'Digital Reconstruction of Cologne Fragments' project.[11]

Very soon after the collapse contact was made with the Fraunhofer Institute for Production and Design Technology (IPK) in Berlin. Members of the Bundestag from Cologne were among those who recalled that this Institute, with its 'puzzle software' for the 'Stasi snippets' ['Stasi-Schnipsel'], had a technology that also seemed suitable for the Cologne fragments.[12] In April 2009, Bertram Nickolay, the head of the department responsible at Fraunhofer IPK, came to Cologne with a number of employees on a first visit – and it was not to be their last, as it quickly became apparent that the expertise developed by the 'Machine Vision' ['Maschinelles Sehen'] department at Fraunhofer IPK could also be used for the Cologne fragments.[13] A feasibility study completed in 2010 identified good opportunities for the digital reconstruction of the ca. three million Cologne fragments. However, it pointed out that, unlike in the case of the destroyed Stasi files, it would be impossible to subdivide the total quantity of snippets into self-contained units (which, in the case of the Stasi records authority, were bags) and that conservation treatment including cleaning would be necessary prior to scanning.

The Historical Archive's cooperation partner in the development of a system for the digital reconstruction of the fragments was MusterFabrik Berlin, a partner company of the Fraunhofer IPK which further develops the existing technology, adapts it to the respective areas of application and was able to help develop work aspects in the area of pure 'Machine Vision'. In 2012, the MusterFabrik was commissioned with a research and development project to demonstrate the technical feasibility of a 'restoration assistance system' in the form of a prototype for the digital and physical restoration of the damaged archive materials of the Historical Archive of the City of Cologne. This was a multi-year project that involved more than just developing software and defining and adapting parameters for the virtual reconstruction of the fragments; rather, it was about the entire process, from 'equipping' the fragments with metadata, through cleaning and smoothing processes, scanning and virtual puzzling, to the storage of the digitised objects and the digitally assembled pages.

This research and development project was uncharted territory for everyone involved as cleaning procedures, the inclusion of existing metadata in the puzzle

11 Hanna Kannengießer and Jan Schneider recently presented an account of these processes and the 'Digital Reconstruction of Cologne Fragments', focussing on the technical aspects: Kannengießer and Schneider 2022; see also Fischer and Späinghaus 2019, 89–93.

12 This technology, developed by Fraunhofer IPK, was used between 2007 and 2016 alongside manual reconstruction of torn State Security ['Stasi'] files from the former GDR.

13 See also Schneider and Nickolay 2015.

algorithms and ultimately also an overall workflow for this process had to be developed from scratch. However, the project partners were able to draw on the existing Fraunhofer reconstruction (Reko) technology as well as on initial experience with the use of scanning technology already commercially available. Another new requirement was the development of a process that would tie up as few resources as possible in terms of archival and restoration specialists; these specialist professionals were to be available to work on those archival units that had been salvaged intact, thus making them available for consultation as quickly as possible.

By the end of the project in 2015, four important development goals were achieved that enabled the start of high-volume operations:

- An entirely newly developed digitisation unit with line scan cameras and its own control software.
- A cleaning workstation where fragile objects can be gently cleaned with cellulose particles.
- A newly developed reconstruction assistance system for the virtual reassembly of the fragments (the actual 'puzzling').
- A procedural outline that embeds these work steps in an overall process and compiles the technical and organisational framework conditions.

In 2015, the development project was successfully completed by proving the functionality of the four components. The task which now remained was to transfer the results and components obtained into an overall project in order to be able to work on a large part of the puzzle fragments in the medium term. To this end the software with its central AI components had to be converted into an intuitively usable system and above all had to be further trained technically with as many fragments as possible.[14] At the same time, the technical expansion of the IT infrastructure had to be implemented and the archiving and restoration processes converted to a high-volume operation. This has been taking place since 2017 as part of the transition from prototype to a high-volume operation.

From the outset the scanning technology has never been a limiting factor as the technical processing can 'keep up with almost any tempo' by scaling the capacities of the IT components. The difficulty lies in the preliminary processes of archival triage ('clearing') and stabilisation (cleaning, smoothing). Since the beginning of the projects solutions have been sought both in-house and together with external

14 AI (or artificial intelligence) is represented in this project by its sub-areas of 'pattern recognition' and 'adaptive learning systems'. This is also where in the case of the Cologne City Archive, the particular strengths of the use of technology for fragment reconstruction lie. However, according to the common definition, this constitutes a 'weak AI' [eine 'schwache KI'], see Bitkom e.V. 2017, 29.

partners to speed up the restoration and conservation processes in particular. The development of efficient processes began even before 2015 with colleagues from the LVR Archive Consulting and Training Centre, where the processing of fragments for the Cologne City Archive was advanced using their own resources.[15]

In the following section a closer look will be taken at some of the processes used to reassemble digitally our fragments with the support of adaptive learning IT systems. Only the virtual reassembling of fragmented archive pages creates the basis for a later physical reconstruction, initially of the fragmented pages and ultimately of entire archival units (files).

To this end the images of the scanned fragments must first be looked at: as a first step these have to be quality-assured anew. Key to this quality assurance is the technically supported visual inspection of whether the torn edges of the fragments are correctly represented on the scanned image.

Metadata that is important in subsequent processes may also be added at this stage. For example, different types of series (e. g. fragments from consecutive pages of a file) can be labelled here. Moreover, information on languages and document types can be added if this information has not already been entered in the salvage entry or during the clearing process.

The additional metadata entered here is part of the 'metadata rucksack' in which the system summarises relevant data for each fragment and makes it usable for the technical processes. What data are these? Firstly, information entered during the early stages of salvaging and recovery is stored in the system for the fragments just like for any other salvage item. Then, the fragment folders, just like all other items, undergoes the so-called salvage registration, i. e. archival items were retrieved or compiled from the salvaged material box by box according to content and physical context; these were described or where possible assigned to a holding or a cataloguing unit. Each fragment now also has this information. In addition, there are special descriptive features entered during the clearing process (e. g. 'sheet music' for all music) as well as those entered after scanning. Needless to say, the storage and transport history of each fragment is also stored in full in the database.

This is the metadata that is contributed with the help of human intelligence. At this point, however, AI comes into play or rather the so-called 'Reko-tags' which are generated independently from the digitised fragments with the help of the software in complex pattern recognition processes. While these cannot be described here in detail, technical processes such as 'polygonal approximation' are standard

15 See e. g. <https://www.lvr.de/de/nav_main/derlvr/presse_1/pressemeldungen/press_report_archiv_13697.jsp> (last accessed 18 October 2023). A total of 182,461 fragments were prepared by the LVR for further processing.

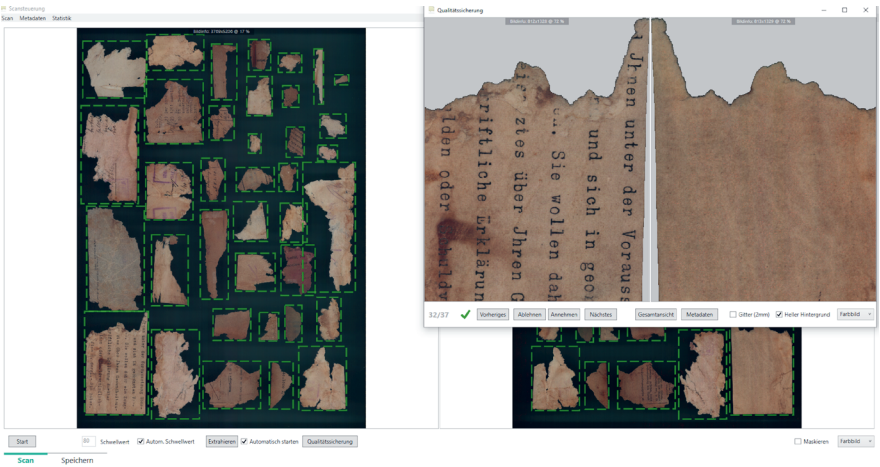

Fig. 2: Quality control on fragments. Photo: MusterFabrik Berlin

tools in digital pattern recognition. Various such tools are combined in the technology created by the MusterFabrik Berlin to form systems that make it possible to capture and describe all newly scanned fragments in such a way that they become available for a puzzle process of a constantly expanding total number of objects (Figure 2).

The technology developed by MusterFabrik Berlin allows fragments to be combined virtually in two steps. In a first step, the search space is reduced on the basis of all the metadata (manually entered along with the automatically extracted data) which means that subsets are formed from the total quantity of scanned fragments according to entered or automatically applied criteria. The automatically generated and stored colour values for paper and typeface play a major role here but other entered and extracted metadata (so-called 'tags') are also used. It is the creation of these subsets that ultimately allows automatic 'matching', i. e. the alignment of each edge with every other edge. Of course, the front and back of both fragments are taken into account during matching, though the automatically recognised outer edges of the leaves are also important. Complex algorithms are used to simplify the geometry of torn edges for preselection and to store section-by-section values for colours for 'puzzling'.

The overall process is actually not that dissimilar to laying out a large jigsaw puzzle where we can observe ourselves carrying out similar steps. We also search for edges, sort by colour spaces and as a last step look at the contours. However, IT-supported pattern recognition is far ahead of us, especially when it comes to processing quantities.

Of course, a comprehensive technical infrastructure is required for this work. After our technical project partner was initially able to set up the technical infrastructure for local operation in our restoration centre, since 2021 the system for fragment reconstruction has been housed securely and in controlled climate conditions in the City of Cologne's computer centre. Many of the racks installed there for housing the digital reconstruction infrastructure are already filled with servers and digital storage systems and further growth may be expected. The application for the 'Digital Reconstruction of Cologne Fragments' (DRKF), by the way, is operated outside the actual municipal network which means that it can be administered and maintained remotely by MusterFabrik Berlin. Individual tasks can be carried out at different locations in Cologne and Berlin.

In the City Archive itself, in addition to the digitisation unit, several so-called 'puzzle workstations' are connected to a 10-gigabit line reserved for the DRKF. At these workstations staff can check and approve the puzzle results in parallel and control the systems, for example for defining and preparing puzzle sets.

But while the DRKF systems can sort the fragments into ever new search spaces at great speed and perform all conceivable puzzle steps they reach their limit elsewhere, as they are unable to read with understanding; it is even more impossible for them to match what has been read with the wealth of experience of the archive staff. For this reason fragment reconstruction was not implemented as an automatic puzzle procedure but as an assistance system. This is because fragments that match well in terms of metadata and torn-edge geometry do not necessarily come from the same page.

An example may illustrate this problem. A volume of a marriage register can contain up to approximately 500 sheets, each of which is printed with the identical sample form for the notarisation of marriages on the front and back; only case-specific data such as names, dates and if applicable document numbers and the like were entered individually by the registrar in each case. If in such a volume many of the pages that are largely identical due to the form used were torn in the same place and with the same pattern due to the collapse (we call this a 'soft series'), the AI can hardly recognise which of the fragments from the series belongs with a given counterpart. This can only be achieved by human intelligence because the human eye immediately recognises an inconsistency in content when reading with understanding. Since this is a common phenomenon the DRKF system does not automatically puzzle together the most probable fragments based on the tags and the torn edges but instead creates a hit list in which pairs of fragments are listed in descending order of probability. The decision to remove a hit – and thus to release for re-puzzling a new fragment formed from the two fragments – is made by the operator. And when the system recognises from the straight outer edges that a (nearly) complete page has been assembled it is presented to an archivist for approval.

Over 320,000 fragments were scanned by the end of 2022. The conservational processing of fragments, the scanning and preparation for puzzling by pattern recognition and the connected processes are ongoing. The support and further development together with the service provider also continue to yield good results but the corona restrictions and the move to the new purpose-built building at Eifelwall have led to a considerable reduction in the number of fragments scanned. However, the transfer of work specifications to new processes has recently been completed so that a moderate increase in throughput may be expected. In addition, as part of the budget planning for the clearing process at the end of 2021, open-ended assistant positions were established which will ensure the continuous operation of the Digital Reconstruction of Cologne Fragments.

This paves the way for the successful continued operation of the project. Of course, this does not mean that intensive development is not continuing in parallel. Thus, a new and more efficient wet cleaning and smoothing process was trialled and introduced in the new restoration workshop at Eifelwall. A considerable proportion of our fragments can now be prepared for scanning in a water bath which means that cleaning and preparatory conservation work such as securing breakage points etc. can be carried out with an improved throughput. The further expansion of the puzzle technology is also on track as is the constant adaptation of the AI to improve the results and the parallel operation of different workstations and in different search spaces.[16]

This also further improves the puzzle results. The still incomplete construction drawing shown in Figure 3 was digitally assembled from a total of 22 fragments in December 2021. Thanks to the continuous addition of new fragments and the performance of new runs with the Reko software, 13 further fragments were identified by the system by August 2022 which further complement the fragmented object.

But, of course, we are looking beyond this. What happens when individual 'puzzled' virtual pages are available again? In our terminology, these are 'formation fragments' that have to be re-combined into a complete file with other 'puzzled' and/or intact pages or parts of the file. A follow-on project for the software-supported Formation of Cologne Fragments is already in preparation with our partner, MusterFabrik Berlin.

Our long-term goal remains that we also want to make it possible to bring together physically the pages 'puzzled' digitally. While this is already being sup-

16 One example of a new development in 2021 is the 'best-fit strategy', which allows the search for fitting fragments from any given match of two or more fragments to continue until straight paper edges on all four sides indicate that a complete sheet of paper has been digitally reconstructed.

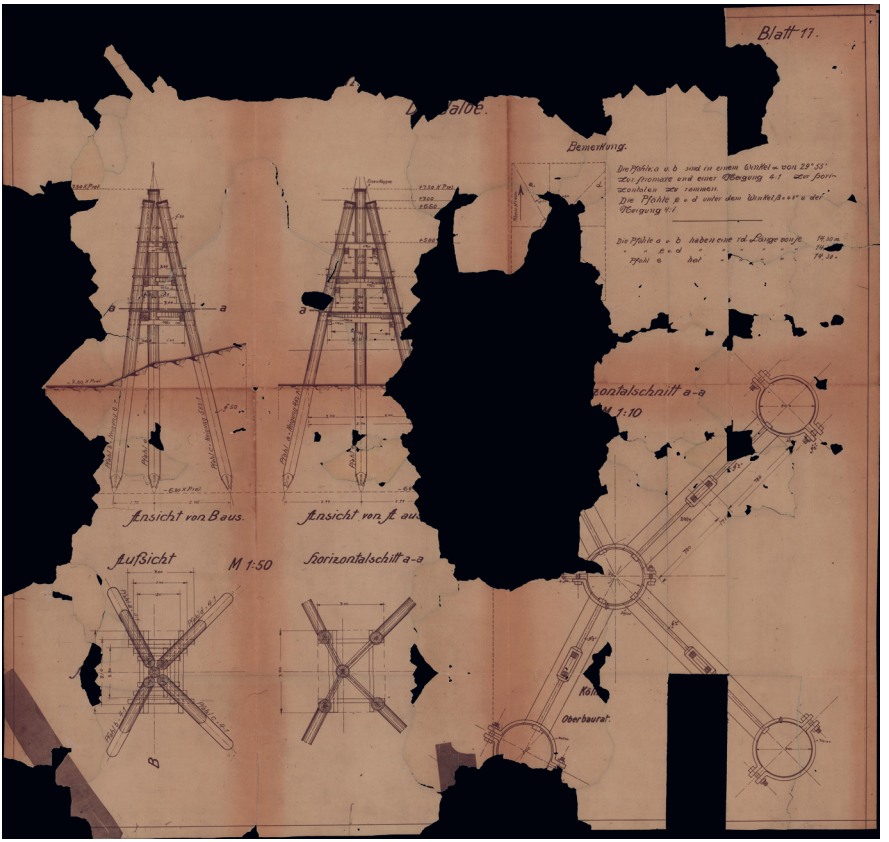

Fig. 3: HAStK, barcode 000497007. Fragment of a construction drawing, virtually assembled from 35 fragments (as of August 2022), MIMG-0006671871_100. Photo: MusterFabrik Berlin

ported by the system, bringing in conservators to reassemble the physical item is still a long way off, especially on a large scale. From a user's point of view, this is only urgent in rare cases. After all, the possibilities offered by IT enable these actually-lost archival documents to be used again – on screen.

References

Bartz, Dietmar (2009). 'Under-Cover-Bericht zum Kölner Archiv (Tag 1): "Wir lesen nicht!"', *taz, die tageszeitung* (15 April 2009). <https://taz.de/!5164673/> (last accessed 18 October 2023).

Bitkom e. V. and Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI) (eds.) (2017). *Künstliche Intelligenz. Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche*

Verantwortung. <https://www.dfki.de/fileadmin/user_upload/import/9744_171012-KI-Gipfelpapier-online.pdf> (last accessed 18 October 2023).

Fischer, Ulrich and Markus Späinghaus (2019). *Geschichte mit Zukunft: 10 Jahre Wiederaufbau des Kölner Stadtarchivs* (Mitteilungen aus dem Stadtarchiv von Köln, Sonderband), Köln: Historisches Archiv.

Fischer, Ulrich and Wolfgang Meyer (2015). 'Die Einsturzfragmente im Kölner Stadtarchiv', in Hanns Peter Neuheuser and Wolfgang Schmitz (eds.). *Fragment und Makulatur. Überlieferungsstörungen und Forschungsbedarf bei Kulturgut in Archiven und Bibliotheken* (Buchwissenschaftliche Beiträge, 91), Wiesbaden: Harrassowitz, 365–376.

Kannengießer, Hanna and Jan Schneider (2022). 'Die digitale Rekonstruktion Kölner Fragmente. Neuland für alle Beteiligten', *Der Archivar* 75: 356–359.

Schneider, Jan and Bertram Nickolay (2015). 'ePuzzler – Automatisierte Virtuelle Rekonstruktion der zerrissenen Stasi-Akten. Technologie und potenzielle Anwendungen', in Hanns Peter Neuheuser and Wolfgang Schmitz (eds.). *Fragment und Makulatur. Überlieferungsstörungen und Forschungsbedarf bei Kulturgut in Archiven und Bibliotheken* (Buchwissenschaftliche Beiträge, 91), Wiesbaden: Harrassowitz, 377–406.