PRZEGLĄD ARCHIWALNY INSTYTUTU PAMIĘCI NARODOWEJ
Tom 8, Warszawa 2015, s. 109–136

Bertram Nickolay, Jan Schneider, Henry Zoberbier*

(Instytut Fraunhofera w Berlinie)

Technologia i możliwe zastosowania zautomatyzowanej wirtualnej rekonstrukcji podartych akt Stasi (ePuzzler)

Jesienią 1989 r. Ministerstwo Bezpieczeństwa Państwowego Niemieckiej Republiki Demokratycznej (Ministerium für Staatssicherheit, MfS lub Stasi) usiłowało w trakcie tajnej operacji pozbyć się ogromnej liczby akt. Pracownicy Stasi podarli ok. 40 mln kartek papieru – na 4 do 30 fragmentów, a niektóre strony nawet drobniej – na 60 i więcej. W ciągu zaledwie kilku tygodni porwano w ten sposób co najmniej 6 kmb akt na ok. 600 mln skrawków, które do dziś są przechowywane w 16 tys. dużych papierowych worków w Urzędzie Pełnomocnika Federalnego do spraw Materiałów Państwowej Służby Bezpieczeństwa NRD (Der Bundesbeauftragte für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik, BStU).

* W Oddziale IPN w Katowicach 19 II 2014 r. odbyła się konferencja pod tytułem „«Uratowana historia» – wykorzystanie technologii informatycznej w procesie rekonstrukcji zniszczonych materiałów archiwalnych". Najnowsze osiągnięcia w dziedzinie rekonstrukcji zniszczonych dokumentów oraz nowe możliwości i korzyści ich zastosowania w archiwach zaprezentowali przedstawiciele: Instytutu Fraunhofera do spraw Systemów Produkcyjnych i Technik Konstrukcyjnych w Berlinie (Fraunhofer Institut für Produktionsanlagen und Konstruktionstechnik; Fraunhofer IPK Berlin) – realizującego projekt wirtualnej rekonstrukcji akt Stasi, oraz firma MFB MusterFabrik Berlin GmbH, która zajmowała się m.in. restauracją zasobu archiwalnego zniszczonego Miejskiego Archiwum Historycznego w Kolonii (Das Historische Archiv der Stadt Köln). Niniejszy artykuł powstał na bazie trzech referatów wygłoszonych na konferencji: dr. Bertrama Nickolaya (Fraunhofer IPK) „Potencjał technologii rekonstrukcji", Jana Schneidera (Fraunhofer IPK) „«ePuzzler» – technologia i zastosowania systemu wirtualnej rekonstrukcji" oraz dr. Marca von der Lindena (MusterFabrik Berlin) „Digitalizacja na skalę masową oraz zabezpieczenie informacji na nośnikach cyfrowych (na przykładach aktualnych projektów)".

Ręczne scalenie tak ogromnej liczby podartych stron możliwe jest tylko częściowo, co więcej, trwałoby niezwykle długo. Dlatego też Instytut Fraunhofera do spraw Systemów Produkcyjnych i Technik Konstrukcyjnych w Berlinie (Fraunhofer Institut für Produktionsanlagen und Konstruktionstechnik; Fraunhofer IPK) w 2007 r. otrzymał od rządu federalnego zadanie kompleksowego opracowania systemu umożliwiającego zautomatyzowaną wirtualną rekonstrukcję podartych dokumentów Stasi. W ramach tzw. fazy pilotażowej za pomocą tego systemu w pierwszej kolejności wirtualnie zrekonstruowanych zostanie ok. 15 mln skrawków stanowiących w przybliżeniu zawartość 400 worków.

W celu realizacji przedsięwzięcia w Instytucie Fraunhofera zaimplementowano program ePuzzler – całkowicie nowy, stworzony do rekonstrukcji, który potrafi połączyć zeskanowane papierowe fragmenty o różnej charakterystyce w kompletne strony. Ponadto oprogramowanie umożliwia ręczną weryfikację i ewentualną korektę wątpliwych lub niejednoznacznych wyników pracy ePuzzlera.

W przeciwieństwie do czysto manualnego sposobu pracy ePuzzler pozwala na obszerną i do tego znacznie szybszą rekonstrukcję zniszczonych dokumentów Stasi. Za pomocą złożonych algorytmów opracowywania obrazu i rozpoznawania wzorów oprogramowanie analizuje dane zdigitalizowanych skrawków. Pasujące elementy odszukiwane są na podstawie takich cech jak kształt, kolor, tekstura, liniowanie i krój czcionki skrawków, a następnie łączone.

1. Wyzwania związane z projektem rekonstrukcji dokumentów Stasi

Podstawowa zasada wirtualnej rekonstrukcji odpowiada metodyce pracy człowieka układającego puzzle. Na podstawie wielu cech podejmuje on decyzję, czy dwie części pasują do siebie, czy też nie. W przypadku większych fragmentów grupuje najpierw podzbiory elementów, które prawdopodobnie pasują do siebie – eliminuje więc przypadek. Analogicznie do sposobu postępowania człowieka ePuzzler analizuje najpierw różne cechy skrawków, takie jak: kontury, kolor papieru, pismo czy liniowanie. Cechy te wykorzystuje się m.in., aby otrzymać liczbę ewentualnych dopasowań na możliwie niskim poziomie poprzez zebranie podobnych skrawków za pomocą zawężenia obszaru poszukiwań w podzbiorach w tak zwane klastry. Właściwa rekonstrukcja odbywa się dopiero w ramach tych klastrów. Jeżeli dwa elementy pasują do siebie, zostają komputerowo połączone, a w dalszej rekonstrukcji uwzględnia się je jako jeden większy fragment. Proces ten określany jest jako *match and merge*. W przypadku dopasowania (*match*) oblicza się stopień zgodności dwóch elementów układanki, a w kolejnym etapie (*merge*) cyfrowo zostają scalone dwa pasujące człony w jedną większą część rekonstrukcji.

Już dzięki uwzględnieniu podczas dopasowywania jedynie przebiegu konturów krawędzi rozdarć – jako cech zgodności – łatwo można z setek skrawków zrekonstruować nieduże części, z których powstanie nawet kilkaset stron. Spełnione jednak muszą zostać następujące warunki:

– przebieg konturów krawędzi rozdarć pasujących par jest w odniesieniu do łącznej liczby opracowywanych skrawków „dostatecznie unikatowy". Tylko wtedy przebieg konturów – porównywalnie do biometrycznego szablonu – może być zastosowany jako selektywne kryterium dopasowania;

– dystrybucja cech, np. koloru skrawków, pisma czy liniowania, wykazuje w odniesieniu do łącznej liczby podlegających opracowaniu skrawków „wystarczającą" liczbę minimów i maksimów. Tylko wówczas można uzyskać efektywne zawężenie obszaru poszukiwań, które jest decydujące dla efektu dopasowania.

Oba warunki w realnych okolicznościach rzadko zostają spełnione, a w przypadku „układanki akt Stasi" – w ogóle. Oznacza to, że system polegający wyłącznie na takich założeniach byłby w praktyce bezużyteczny. Dowodzą tego następujące kwestie:

1. Rozrywanie jest formą niszczenia.

W zależności od rodzaju papieru rozrywanie może prowadzić do trwałego zniszczenia struktur włókien papieru w obszarze krawędzi rozdarcia. Kilka rodzajów papieru w czasie jego wykorzystywania, transportu oraz rozrywania wykazuje tendencję do „strzępienia się". Dochodzi tutaj do utraty materiału, skutkiem czego są „naturalne" luki w parach krawędzi rozdarć podlegających rekonstrukcji.

2. Rozdarty papier bardzo rzadko można idealnie dopasować.

Już w przypadku przeciętnie grubego papieru (od ok. 70g/m²) na krawędziach rozdarć pojawia się tzw. ścinanie. W jego obszarach pasujące skrawki nie dają się złożyć *per se* na styk, lecz muszą zostać nałożone miejscowo.

3. Pary krawędzi ścierania rzadko są unikatowe.

Dokumenty często rozrywa się w całości. W celu przyspieszenia ręcznego procesu darcia poszczególnych kart są one układane w stos (niekiedy kilka kart danego dokumentu zostało spiętych) i dopiero wówczas rozrywane jako całość. W efekcie powstaje bardzo wiele podobnych brzegów rozdarcia, które podczas stosowania konturu jako kryterium dopasowania dają niejednoznaczne wyniki (por. też ilustrację nr 1).

4. Rzeczywiste dokumenty mogą w zależności od danej karty mieć bardzo różną charakterystykę.

W praktyce poszczególne dokumenty w przypadku zarówno karty, jak i strony (przedniej i tylnej) mogą wykazywać duże zróżnicowanie w obrębie ich cech. Teksty napisane na maszynie mogą np. mieć odręczne notatki sporządzone przez kilka osób, ponadto jedna kartka może być częściowo pożółkła lub wyblakła. Powoduje to, że poszczególne pasujące skrawki mogą wykazywać odmienne tekstury, kolory pisma czy barwy papieru. Generalnie: różnorodność cech skrawków nie może być zatem stosowana jako kryterium wyłączające przy zawężaniu obszaru poszukiwań.

5. Wszystkie karty rzeczywistego dokumentu mogą mieć bardzo jednolitą charakterystykę.

Często (w przeciwieństwie do poprzedniego punktu) różnorodność ogólnych cech wielu tysięcy skrawków może być niewielka, np. gdy podarty został kilkusetstronicowy dokument napisany na maszynie. Rozrywanie dotyczy z reguły kilku-, kilkunastu stron jednocześnie. Dla takiej liczby skrawków nie jest możliwe zawężenie obszaru poszukiwań. Rezultatem są maksymalnie „opisane" i „puste" skrawki w obu klasterach.

Powyższy stan rzeczy, występujący niemalże we wszystkich tego typu przedsięwzięciach, należy w kontekście dokumentów Stasi dodatkowo uzupełnić o następujące punkty:

6. Wariancja liczby wszystkich skrawków jest niezwykle duża.

Akta zakładano w zasadzie od utworzenia Stasi w 1950 r. aż do jej rozwiązania w 1990 r. Część dokumentów dotyczy jeszcze okresu III Rzeszy, a nawet wcześniejszego. Tworzyło je wiele tysięcy osób, Stasi bowiem przez cały czas swojego funkcjonowa-

nia zatrudniała około 250 tys. etatowych i więcej niż 600 tys. nieoficjalnych pracowników. Podarte zostały zatem akta wszelkiego typu i różnej charakterystyki.

7. Układanka jest niekompletna.

Ze względu na okoliczności przekazania jest wysoce prawdopodobne, że w workach brakuje części poszczególnych stron. Nie jest niczym dziwnym, że w zawirowaniach panujących w czasie zmian ustrojowych nie zważano na to, aby wszystkie skrawki podartych dokumentów trafiły do właściwych pojemników. Niekompletność przekazanych worków znacznie podnosi złożoność procesu rekonstrukcji.

8. Krawędzie rozdarć poszczególnych zestawień par skrawków pasują tylko częściowo.

Podczas jednoczesnego rozrywania całych dokumentów liczących wiele stron – tak jak opisano to w punkcie trzecim – często powstają, w szczególności w przypadku stron znajdujących się wewnątrz pliku, niewielkie skrawki o średnicy kilku milimetrów. Nawet jeżeli te najdrobniejsze kawałki nie zaginęły, zgodnie z punktem siódmym można je zdigitalizować i zrekonstruować tylko przy użyciu niewspółmiernie dużych nakładów. Dlatego też elementy te są odkładane na bok i ewentualnie przekazane do osobnego opracowania. Powoduje to, że krawędzie rozdarcia pasujących skrawków mogą wykazywać znacznie większe braki, niż opisano to w punkcie pierwszym.

9. Kombinatoryczny wkład pracy w rekonstrukcję jest niezmiernie duży.

Biorąc pod uwagę okoliczności przekazania, wiadomo (na szczęście), że nie chodzi jedynie o układankę składającą się z 600 mln części. Chociaż podarte dokumenty zabezpieczono w różnych miejscach – w kontekście archiwalnym określanych jako proweniencja – nie należy zakładać, że skrawki tych samych materiałów można odnaleźć w workach różnej proweniencji. Próby losowe pokazały, że w ramach proweniencji dokumenty rozdzierano i umieszczano celowo w różnych workach. Tak więc podczas procesu odtwarzania należy skupić się nie na zawartości poszczególnych worków jako początkowej liczbie elementów do ułożenia, ale na zawartości wszystkich worków danej proweniencji. W najgorszym przypadku jedna proweniencja obejmuje kilkaset worków.

10. Błędne kroki rekonstrukcyjne multiplikują się poprzez proces, zwiększają nakład kombinatoryczny, a od pewnego rzędu wielkości skrawków są niemożliwe do opracowania.

W trakcie opracowywania i implementacji programu ePuzzler uwzględniono warunki ramowe i okoliczności określone w punktach od 1 do 10. Ze względu na czas i wysokie koszty uwagą objęto także wytyczne dotyczące przydatności w praktyce oraz wydajności.

W wielu specjalistycznych artykułach (niezbyt licznych na ten temat) czyni się idealne założenia i opisuje sposoby rozwiązań, które w określonych warunkach dają nadzwyczaj dobre wyniki w rozpoznawaniu, a także nie najgorsze wskaźniki rekonstrukcji. Wymienia się np. następujące – często domyślnie założone – warunki „brzegowe": kompletność skrawka – puzzla; rekonstrukcja niewielu stron; wykluczające krawędzie rozdarć (częściowo wygenerowane tylko syntetycznie); biały papier z czarnym pismem (np. nowoczesny wydruk laserowy); wiedza *a priori* o *layoucie* lub o formacie dokumentów; wiedza *a priori* o orientacji skrawków. Teorie te są interesujące z akademickiego punktu widzenia, ale niestety nieprzydatne do rozwiązania realnych zadań (od pewnego rzędu wielkości). Dlatego też przy rekonstrukcji dokumentów Stasi można je było wziąć pod uwagę wyłącznie warunkowo. Podczas opracowywania całościowej logiki ePuzzlera,

Ilustracja nr 1. Przykład dokumentów rozdartych w całości z bardzo podobnymi krawędziami rozdarcia. Kilka stron dokumentu rozerwanego za jednym razem prowadzi z reguły do wieloznacznych propozycji rekonstrukcyjnych. Z lewej strony i w środku: propozycje dopasowania skrawka z tekstem DDR jest niewłaściwe. Z prawej strony: połączenie skrawka z tekstem DDR jest poprawne

Źródło: Fraunhofer IPK, 2014

w szczególności przygotowywania różnorodnych modułów badania obrazu i rozpoznawania wzorów, na których bazuje ePuzzler-workflow, od początku stosowano się do poniższych zasad:

1. Nie istnieje *a priori* żadna wiedza dotycząca formatu stron podlegających rekonstrukcji.

2. Nie istnieje *a priori* żadna wiedza dotycząca zawartości stron podlegających scaleniu.

3. Nie istnieje algorytmika, która sprawia, że dowolne teksty stają się czytelne dla komputera.

4. Nie istnieje żadna ogólnie obowiązująca ilość obrazowych cech pozwalająca na podjęcie jednoznacznej decyzji, czy „dopasowanie jest poprawne/nie jest poprawne".

5. Nie może istnieć górna granica systemowa do jednocześnie opracowywanych skrawków.

6. Liczbę elementów układanki należy utrzymywać na możliwie niskim poziomie.

Dzięki implementacji inteligentnych technik opracowania obrazu oraz metod filtracji przebiegu pracy (*workflow*) stało się możliwe opracowywanie pofragmentowanych dokumentów Stasi. Nie poczyniono przy tym żadnych założeń *a priori* dalekich od praktyki. Dostępne cechy ekstrahuje się raczej każdorazowo w odniesieniu do skrawka. To znaczy, że zasadniczo nie szuka się określonych cech, lecz bez uprzedniej znajomości charakterystyki danego kawałka odszukiwane są możliwie wszystkie cechy danego elementu. Stawia to wyraźnie wyższe wymagania algorytmice oraz wymaga mocnych

i czasowo efektywnych metod przetwarzania obrazu i rozpoznawania wzorów. Moduły opracowane w ramach projektu pracują w trybie Puzzle-workflow, do tego sukcesywnie wg zasady „najpierw *low-level* (np. kolor papieru), potem *high-level* (np. frekwencja wierszy)", aż do wyselekcjonowania coraz bardziej prawdopodobnych fragmentów do połączenia. Dopiero to ciągłe wykluczanie wątpliwych elementów prowadzi do znacznej redukcji liczby skrawków w całości układanki, tj. do zachowania szóstej zasady, i umożliwia złożenie pofragmentowanych dokumentów Stasi. Właściwy tryb pracy jest realizowany dzięki wrażliwemu na kontekst oprogramowaniu strukturalnemu, które zostało ostatnio opracowane przez Instytut Fraunhofera i ukierunkowane na maksymalny stopień paralelizacji.

2. Sposób funkcjonowania ePuzzlera

Opracowany przez Instytut Fraunhofera system ePuzzler stanowi szkielet zautomatyzowanej wirtualnej rekonstrukcji pofragmentowanych dokumentów Stasi. Zautomatyzowany oznacza w tym przypadku, że odbywa się ona we współpracy z jednym lub kilkoma operatorami. Proces rekonstrukcji nie jest więc przeprowadzany w pełni automatycznie, bez ingerencji człowieka. Jednakże wszystkie kroki procesu wymagające manualnego działania są przygotowywane automatycznie przez ePuzzlera. A więc operator, oszczędzając czas, może wykonywać po kolei odpowiednie czynności, zgodnie z określonym schematem. Zastosowane przy tym narzędzia są ściśle zintegrowane z ePuzzlerem i stanowią istotny element całego systemu, wykraczający poza właściwe „składanie puzzli".

Poniżej krótko opisano etapy procesu zautomatyzowanej wirtualnej rekonstrukcji.

1. Wprowadzenie do systemu śledzenia skanera informacji przekazanych przez pracowników Urzędu Pełnomocnika Federalnego ds. Materiałów Stasi.

Przed digitalizacją skrawki w trakcie tzw. wstępnego sortowania są ręcznie przygotowywane przez pracowników Urzędu Pełnomocnika Federalnego ds. Materiałów Stasi. Podczas tych czynności kawałki są sortowane według transportu i jednostki opracowania (*TVe – Transport und Verarbeitungseinheit*), zgodnie z kryteriami obowiązującymi w Urzędzie Pełnomocnika Federalnego, oraz pakowane do kartonów. W czasie tego procesu usuwane są np. spinacze, przyporządkowane zostają koperty z oddzielonym materiałem, a elementy, które nie podlegają wirtualnej rekonstrukcji, są pakowane do osobnych kartonów. Wszystkie koperty, kartony itp. otrzymują numer odpowiadający właściwemu transportowi i jednostce opracowania TVe (numer transportu i opracowania), tj. TVN (*TVN – Transport- und Verarbeitungsnummer*), po czym zostaje on na nich umieszczony.

Ponadto wszystkie skrawki są dzielone przez pracowników Urzędu Pełnomocnika Federalnego ds. Materiałów Stasi na różne kategorie i odpowiednio oznaczane, m.in. jako towar A i towar B, a także znacznik K i znacznik S. Oznaczenia „towar A i B" to efekt oceny zawartości dokonanej przez Urząd Pełnomocnika Federalnego, gdzie towar B zawiera najprawdopodobniej treści nieistotne, które zapewne po rekonstrukcji będzie można usunąć. Znacznik K otrzymują małe fragmenty, a znacznik S – skrawki należące do akt tej samej sprawy (ewentualnie spięte razem). Wszystkie oznaczenia muszą zostać uwzględnione w trakcie digitalizacji, a także w następującym po niej procesie rekonstrukcji.

2. Digitalizacja.

Podczas digitalizacji odbywa się przeniesienie fizycznych elementów do świata wirtualnej rekonstrukcji. Poza właściwym utworzeniem obrazu skrawka przeniesione także zostają – w formie metadanych jako jeden fragment układanki – wszystkie znane informacje niezbędne do dalszego opracowania.

Każdy fragment skanowany jest w tym celu z obu stron, w wyniku czego ma on w systemie dwa obrazy. Powstała w taki sposób para skrawków otrzymuje bieżące oznaczenie, które zawiera numer transportu i opracowania TVN, dane kartonu oraz zapisane wcześniej cechy kategorii.

3. Import zdigitalizowanych skrawków.

Obrazy skrawków danej jednostki opracowania, tj. z reguły wszystkie skrawki z jednego numeru transportu i opracowania TVN, przed przystąpieniem do układania (puzzlowania) muszą zostać wgrane do systemu ePuzzler. W celu przeprowadzenia importu należy w odpowiedniej kolejności wykonać następujące etapy procesu:

a) zapewnienie jakości zdigitalizowanych obrazów skrawków, ewentualnie wraz z przygotowaniem błędnych kopii cyfrowych oraz wysortowaniem małych skrawków nieprzydatnych do opracowania;

b) automatyczne zapisanie obrazów skrawków w bazie danych;

c) ewentualne ręczne dzielenie obrazów skrawków danej jednostki opracowania na podzbiory, które charakteryzują się różnym priorytetem przy układaniu fragmentów, np. niski mają niewypełnione druki formalne, niezapisane karty kalendarzy na biurko itp.[1]

4. Układanie (*puzzle*).

Obrazy skrawków danej jednostki opracowania układane są wg strategii drzewa binarnego, zgodnie z zasadą „każdy przeciw każdemu". Program automatycznie oblicza prawdopodobne dopasowania elementów i/lub częściowe rekonstrukcje („kandydaci"). Wszystkie fragmenty wykazujące dostatecznie wiele cech wspólnych (np. przebieg konturów, pismo, liniowanie itp.) są automatycznie składane. „Kandydaci" wykazujący słabe lub wzajemnie wykluczające się cechy (np. przebieg konturów pasuje, ale nie ma kontynuacji pisma) zostają wyświetleni na monitorach komputerów operatorów Urzędu Pełnomocnika Federalnego ds. Materiałów Stasi („propozycje"). Wszystkie propozycje przyjęte przez pracowników Urzędu Pełnomocnika Federalnego zostają razem poskładane, natomiast pozostałe – odrzucone. Tym samym opracowanie danej jednostki przypomina grę w ping-ponga, wg automatycznych kroków ePuzzlera i z odręcznymi działaniami operatora.

5. Zapewnienie jakości (QS) (*QS – Qualitätssicherung*).

Na koniec opracowania danej jednostki wszystkie rekonstrukcje przeprowadzone w poprzednim etapie procesu muszą zostać poddane manualnemu procesowi zapewnienia jakości. Pracownicy Urzędu Pełnomocnika Federalnego w dowolnej kolejności na swoich komputerach przeprowadzają następujące kroki procesu:

a) potwierdzenie poprawnej całkowitej rekonstrukcji;

b) usunięcie błędnie ułożonych skrawków w częściowej lub całkowitej rekonstrukcji;

[1] Całkowicie puste skrawki mogą zostać wysortowane przez ePuzzlera automatycznie, ewentualnie – otrzymać niski priorytet. Ponieważ niewypełnione druki w sensie obrazowym nie są puste, ustalenie priorytetu tego typu musi odbywać się ręcznie.

c) zmiana statusu częściowej rekonstrukcji na całkowitą[2];

d) ewentualne drobne ustawienia skrawków zarówno w rekonstrukcjach całkowitych, jak i w częściowych zmienionych na całkowite.

6. Przygotowanie wyników rekonstrukcji (PDF/A i wydruk).

Po zakończeniu etapu zapewnienia jakości wszystkie rekonstrukcje całkowite oraz częściowe zmienione na całkowite danej jednostki opracowania konwertowane są automatycznie przez ePuzzlera do formatu archiwizacji PDF/A. Dla każdej rekonstrukcji generowane są dwa pliki PDF/A: jeden dla awersu, drugi dla rewersu. Wszystkie pliki PDF/A oraz pozostałe parametry danej jednostki opracowania przekazywane są Urzędowi Pełnomocnika Federalnego ds. Materiałów Stasi w formie zapisu liniowego na taśmach LTO lub na DVD.

Wszystkie pliki PDF/A, które nie są puste, zostają ponadto wydrukowane, a następnie przekazane do Urzędu. Wydruk zostaje wykonany za pomocą narzędzi zintegrowanych z systemem ePuzzler.

W następnej części artykułu opisano szczegółowo trzy etapy procesu: digitalizację, układanie (wraz z zapewnieniem jakości) oraz przygotowanie wyników rekonstrukcji.

2.1. Digitalizacja

Cały przebieg digitalizacji składa się z czterech uzależnionych od siebie faz, z których każda może zostać przeprowadzona w odrębnym czasie (por. ilustracja nr 2). Taki sposób postępowania umożliwia efektywny przebieg całego procesu.

Ilustracja nr 2. Etapy całego przebiegu digitalizacji

Źródło: arvato AG[3]

Fragmenty dokumentów Urzędu Pełnomocnika Federalnego różnią się w kilku istotnych punktach od tych opracowywanych w ramach tradycyjnej digitalizacji dokumentów. Z reguły dotyczy to skrawków pochodzących ze zwykłego papieru. W pojedynczych przypadkach fragmenty są z papieru o większej gramaturze (np. kartonowe karty map) lub z bardzo cienkiego papieru (kalka do maszyny do pisania). W związku z celowym i gwałtownym niszczeniem dokumentów przez Stasi zachowane fragmenty są skrawkami niemającymi określonego formatu czy konturu. W zależności od stopnia zniszczenia

[2] Pełnomocnikowi federalnemu przekazywane są jedynie rekonstrukcje całkowite oraz częściowe zmienione na całkowite. Jeżeli rekonstrukcja częściowa, w której najprawdopodobniej brakuje jedynie pustych skrawków, nie może przejść na poziom statusu rekonstrukcji całkowitej, Urząd Pełnomocnika Federalnego nie uzyska dostępu do jej zawartości.

[3] Arvato AG to spółka powstała 1 VII 1999 r. w Niemczech, która zajmuje się outsourcingiem i stanowi dział międzynarodowego koncernu Bertelsmann SE & Co.KGaA. Jej nazwa to akronim od słów: Ars für Variation für Technik und für Organisation.

możliwy jest zarówno wszelki kształt i w przybliżeniu każda wielkość oraz dowolna kolejność elementów – od mniejszych niż 2 × 2 cm do formatu A3. Dodatkowe wymieszanie fragmentów doprowadziło do tego, że w workach, w których przechowywane są skrawki, panuje tylko częściowy porządek – tak więc obok bardzo małego kawałka może pojawić się element rozmiaru A5.

Ilustracja nr 3. Fragmenty dokumentów Urzędu Pełnomocnika Federalnego ds. Materiałów Stasi

Źródło: Urząd Pełnomocnika Federalnego ds. Materiałów Stasi

Kawałki przekazane przez Urząd Pełnomocnika Federalnego ze względu na ich przeznaczenie (zniszczenie) oraz wskutek przechowywania w papierowych workach mogą znajdować się w różnym stanie – być pozaginane, naderwane, postrzępione, poobcinane lub uszkodzone w inny sposób. Pierwotny stan fizyczny fragmentów stanowi tym samym duże wyzwanie w zakresie przygotowania i dalszej digitalizacji.

W ramach projektu dotyczącego materiałów Stasi obecnie do digitalizacji stosuje się wielkoformatowy skaner duplex, który jest w stanie skopiować dokumenty zadrukowane z obu stron, o szerokości do 36 cali (915 mm). Urządzenie, pierwotnie przeznaczone do digitalizacji obu stron gazet, zostało zmodyfikowane na podstawie specjalnych wymagań w zakresie digitalizacji fragmentów dokumentów. W celu digitalizacji skrawków papieru kierowano się m.in. zasadą nośników obiektów. Postępowanie to odpowiada znanej metodzie z mikroskopii, polegającej na umieszczaniu małych obiektów na przeźroczystych nośnikach o znanej wielkości w celu lepszego posługiwania się nimi. Chodzi tutaj o opakowania foliowe o wymiarach 660 × 597 mm, w których umieszcza się skrawki przeznaczone do digitalizacji. Folia wprowadzana jest do skanera, który ją pobiera, blokuje, a następnie automatycznie przeciąga i wyrzuca z tyłu. Poszczególne warstwy folii mają grubość ok. 165 µm, są antystatyczne z obu stron i minimalnie odbijają światło, wykazują znikome zużycie abrazyjne i wszystkie są odporne na zgięcia.

Ze skanerem zintegrowane są dwa aparaty fotograficzne. Taka budowa umożliwia digitalizację zgodną z orientacją i jednakowym pokryciem zarówno awersu, jak i rewersu skrawków znajdujących się w trakcie kopiowania w foliach.

Dzięki zastosowaniu folii kurz i zanieczyszczenia, które znajdują się jeszcze na skrawkach, pozostają wewnątrz i nie brudzą części urządzenia, np. delikatnych elementów optyki. Folię w razie konieczności można wyczyścić, a w ostateczności – wymienić. Ponadto ułożenie w nich skrawków zapobiega ich bezpośredniemu obciążeniu w trakcie procesu.

Tym samym chodzi o wyjątkowo ostrożny sposób digitalizacji, za pomocą którego można skanować także bardzo podarte i łamliwe fragmenty. Ścieranie papieru praktycznie nie występuje, ponieważ skrawki przekładane są jedynie z kartonu archiwizacyjnego do folii i ponownie do kartonu.

2.1.1. Przygotowanie

Fragmenty przeznaczone do skanowania przekazywane są przez pracowników Urzędu Pełnomocnika Federalnego w kartonach przeznaczonych do archiwizacji, wykonanych z tektury mikrofalistej, bezkwasowej i odpornej na starzenie. Każdy z przygotowanych kartonów ma w momencie przekazania widoczną etykietę i zostaje zarejestrowany w systemie śledzenia skanera.

Ilustracja nr 4. Przykład wypełnionych kartonów archiwizacyjnych. Z prawej strony seria skrawków spiętych razem w jednej kopercie

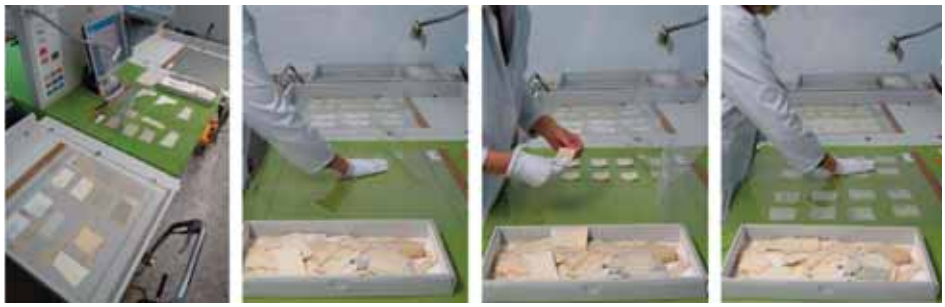
Źródło: arvato AG

W ramach przygotowań karton ze skrawkami przeznaczonymi do zeskanowania zostaje pobrany z magazynu tymczasowego. Skaner odczytuje z etykiety z kodem paskowym proweniencję i numer kartonu, a system śledzenia zapisuje status „Przygotowanie".

W pierwszej kolejności każdy fragment układa się osobno, a te mocno pofałdowane poddaje na wstępie delikatnemu wygładzeniu. Ostrożnie usuwane są również obce ciała (spinacze, zszywki, materiał, który nie jest papierowy, jak np. taśmy filmowe[4]) oraz zagięcia, tak aby można było włożyć do folii same skrawki. Ponadto między poszczególnymi fragmentami musi pozostać odpowiednia ilość wolnego miejsca, żeby w trakcie digitalizacji nie zostały omyłkowo rozpoznane jako pojedynczy element.

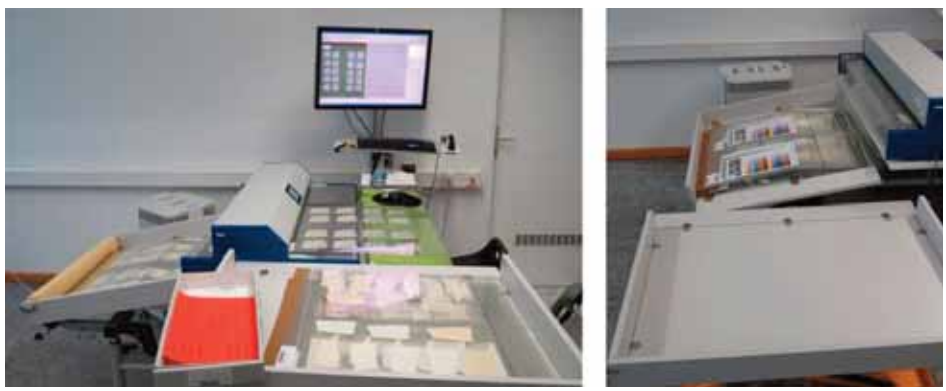
Aby w procesie rekonstrukcji można było zastosować wszystkie oznaczenia materiałów, jakie zostały nadane im przez pracowników Urzędu Pełnomocnika Federalnego, stosuje się kody paskowe do znakowania skrawków. W czasie przygotowań kody paskowe wkładane są do folii i odczytywane przez skaner w trakcie digitalizacji. W zależności od treści zapisanej pod danym kodem paskowym odpowiednie informacje są systematycznie rejestrowane i łączone z właściwym skrawkiem.

[4] Materiał nienadający się do zeskanowania jest zwracany Urzędowi Pełnomocnika Federalnego wraz z informacją o kartonie archiwalnym, z którego został pobrany.



Ilustracja nr 5. Przygotowanie folii. Od lewej do prawej: miejsce przygotowań, oczyszczanie folii, układanie skrawków, przykrywanie folią

Źródło: arvato AG



Ilustracja nr 6. Stanowisko pracy w procesie skanowania. Z lewej: w trakcie skanowania; z prawej: pakowanie już zdigitalizowanych folii

Źródło: arvato AG

Ilustracja nr 7. Moduł skanowania – własny projekt Instytutu Fraunhofera

Źródło: Fraunhofer IPK, 2014

Ilustracja pokazuje proces umieszczania elementów w folii. Najpierw czyści się ją od wewnątrz. Następnie z kartonu poszczególne fragmenty są wyjmowane, oceniane, a następnie przygotowywane tak, aby nadawały się do zeskanowania, dopiero po tych czynnościach układane są na folii. Po kompletnym zapełnieniu folii zostaje ona wyczyszczona od zewnętrznej strony i odłożona wraz z pozostałymi.

Ponieważ przedmiotem przygotowań jest papier, w trakcie tych czynności powstaje dużo kurzu. W celu utrzymania czystości folii na jak najwyższym poziomie, a przede wszystkim uniknięcia błędnych kopii cyfrowych w wyniku obecności obcych cząsteczek zarówno wewnątrz, jak i na folii (kurz z papieru, włókna, włosy, brud itp.), pracownicy w trakcie procesu noszą rękawiczki oraz fartuchy robocze, które zapobiegają przenoszeniu odcisków palców oraz włókien z odzieży na folie. Ponadto zainstalowano filtry oczyszczające powietrze.

2.1.2. Skanowanie

Dla każdego zdigitalizowanego kartonu w systemie śledzenia generowane jest własne zadanie skanowania, które oznaczane jest wg proweniencji i numeru kartonu. Kod paskowy kartonu przeznaczonego do skopiowania zostaje odczytany i w systemie śledzenia pojawia się status „Skanowanie". Wszystkie folie danego kartonu archiwizacyjnego są digitalizowane po kolei. Przeciągane są pojedynczo przez skaner, digitalizowane z obu stron, a następnie wyrzucane do specjalnego pojemnika przechwytującego.

Ilustracja ukazuje stanowisko pracy w czasie skanowania. Na pierwszym planie widoczny jest wózek transportowy z foliami przeznaczonymi do skanowania oraz kartonem archiwizacyjnym, na drugim planie – skaner. Na monitorze przy ścianie można dostrzec pierwszy obraz prezentujący zdigitalizowaną folię.

Po wykonaniu digitalizacji wszystkich folii z danego kartonu „zadanie skanowania zostaje zakończone". Każdy obraz skrawka otrzymuje unikatową (z kolejnym numerem) nazwę pliku, w którym zakodowane są m.in. numer worka i kartonu, a także rodzaj materiału i pozostałe cechy. Numeracja fragmentów znajdujących się w folii odbywa się w kolejności, według wierszy, od lewej górnej strony do prawej dolnej. Obrazy skrawków z zakończonego zadania skanowania zostają ułożone wg tymczasowego spisu w celu wykonania kontroli jakości.

2.1.3. Zapewnienie jakości

W celu zagwarantowania prawidłowego przebiegu czynności został zdefiniowany wieloetapowy proces zapewnienia jakości. Każda kopia cyfrowa musi zostać uważnie oceniona, aby do etapu rekonstrukcji nie przekazano mylnych obrazów. Za potencjalne błędy uważa się na przykład „ośle uszy" (zagięte rogi kartek), obce cząsteczki, jak brud czy włókna na obrazach, a także zakłócenia obrazu, np. cienie i barwne zniekształcenia. Ponadto bezpośrednio po digitalizacji folii operator skanera ma za zadanie skontrolować, czy uzyskane obrazy są kompletne. Sprawdzane jest, czy wszystkie umieszczone w folii skrawki zostały poprawnie odseparowane oraz czy można wygenerować odpowiednie pojedyncze obrazy.

Ponieważ proces rekonstrukcji uzależniony jest w znacznej mierze od jak najwierniejszego odwzorowania, tj. od najwyższej jakości i braku błędów w obrazach, w celu zapewnienia jego odpowiedniej jakości stworzono specjalne stanowisko pracy. Dodatkowa osoba przed ostateczną akceptacją uzyskanych skanów ponownie kontroluje każdy pojedynczy obraz. Ta dodatkowa weryfikacja chroni przed zwykłymi ludzkimi błędami.

2.1.4. *Postprocessing*

W ramach *postprocessingu* fragmenty dokumentów zostają usunięte z zeskanowanych folii i ponownie schowane w kartonach archiwizacyjnych. Dzięki ułożeniu wszystkich skrawków ze znacznikami w odpowiednich kopertach zostaje w kartonach przywrócony poprzedni układ. Zarówno stanowiska pracy przygotowania, jak i *postprocessingu* są tak samo wyposażone i praca na nich przebiega wg tej samej systematyki.

Wykorzystana folia jest sprawdzana, czy nadaje się do ponownego użytku. Ponieważ na skrawki negatywnie oddziałują głębokie zarysowania folii, można do skanowania używać ich tylko ograniczoną liczbę razy. Folie z widocznymi brakami jakościowymi zostają wycofane. Ponadto każda z nich ma własny kod paskowy, dzięki któremu można ustalić liczbę skanowań. Po osiągnięciu maksymalnej liczby użyć automatycznie wyświetlany jest komunikat, że folię należy usunąć.

2.1.5. Perspektywy

Zastosowanie folii w procesie skanowania okazało się efektywne. Zarówno uszkodzone i rozdarte fragmenty, jak i bardzo małe kawałki po umieszczeniu ich w folii doskonale nadają się do digitalizacji. Utrudnieniem jest fakt, że trzeba to robić ręcznie. Otwieranie i zamykanie oraz układanie w nich fragmentów jest niestety bardzo czasochłonne. Pozytywnym aspektem jest jednak to, że użycie folii umożliwiło oddzielenie etapu przygotowań fragmentów dokumentów od właściwego procesu skanowania. Na osobnych stanowiskach pracy można zapełniać folie, a po ich digitalizacji opróżniać, a w tym czasie digitalizować kolejny karton. To samo dotyczy stanowiska *postprocessingu*. Dzięki usystematyzowaniu przebiegu pracy zaoszczędzono cenny czas.

Przebieg procesu digitalizacji opracowany w ramach projektu pilotażowego umożliwia zeskanowanie skrawków dokumentów Stasi zgodnie z minimalnymi wymogami. Wysokie wymagania jakościowe są spełnione w wystarczającym stopniu dzięki wieloetapowemu zapewnieniu jakości. Jednakże zastosowane rozwiązania są niezwykle czasochłonne. Z tego też względu z inicjatywy Instytutu Fraunhofera rozpoczęto prace nad koncepcją i prototypowym opracowaniem nowej technologii skanowania, która nadawałaby się do masowych zastosowań oraz odpowiadałaby wysokim wymaganiom zautomatyzowanej wirtualnej rekonstrukcji. Celem jest możliwie jak najkrótsza droga digitalizacji.

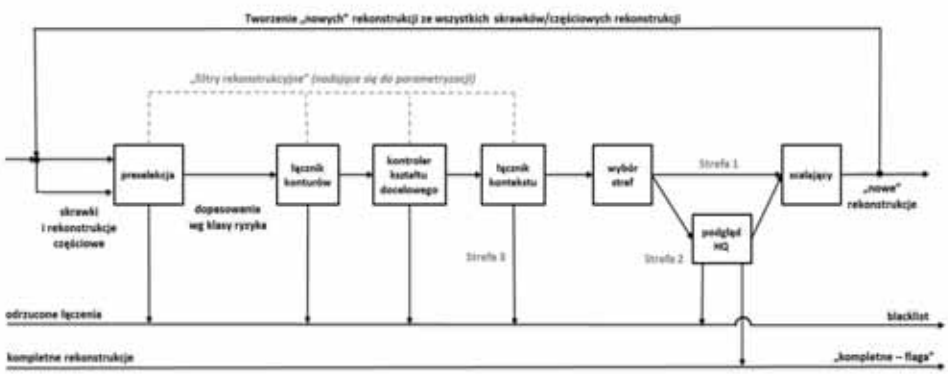
Moduł skanowania, „serce" drogi digitalizacji, został już zakupiony ze środków Instytutu Fraunhofera, niezależnych od projektu pilotażowego, a także zmontowany (ilustracja nr 7). Funkcjonalność tego prototypu obejmuje obustronny zapis obrazu z miejscowymi wahaniami rozdzielczości w niskich zakresach (w promilach) oraz wahania barw poniżej poziomu dostrzegalnego przez człowieka.

W przyszłości planowane jest innowacyjne opracowanie poszczególnych komponentów i modułów, które umożliwiłyby jak najbardziej zautomatyzowane wprowadzanie skrawków do nośników skanera. Konieczne będzie zastosowanie stabilnych nośników skanera mających optyczne właściwości przystosowane do opracowywania obrazów.

2.2. Układanie (*puzzling*) i zapewnienie jakości

Wszystkie skrawki, jak już wspomniano, oraz rekonstrukcje częściowe poddawane są określonemu przebiegowi pracy (RECO-workflow), realizowanemu przez oprogramowanie strukturalne wrażliwe na kontekst. Jednym z głównych zadań RECO-workflow

jest utrzymanie kombinatorycznych nakładów zasadniczego układania (dopasowanie 1:1) na tak niskim poziomie, jak to możliwe. W tym celu dla wszystkich skrawków i częściowych rekonstrukcji danej liczby opracowań stosuje się wiele filtrów (RECO-filter).



Ilustracja nr 8. Przebieg pracy (*workflow*) – wycinek: filtry RECO i klasyfikacja strefowa

Źródło: Fraunhofer IPK, 2014

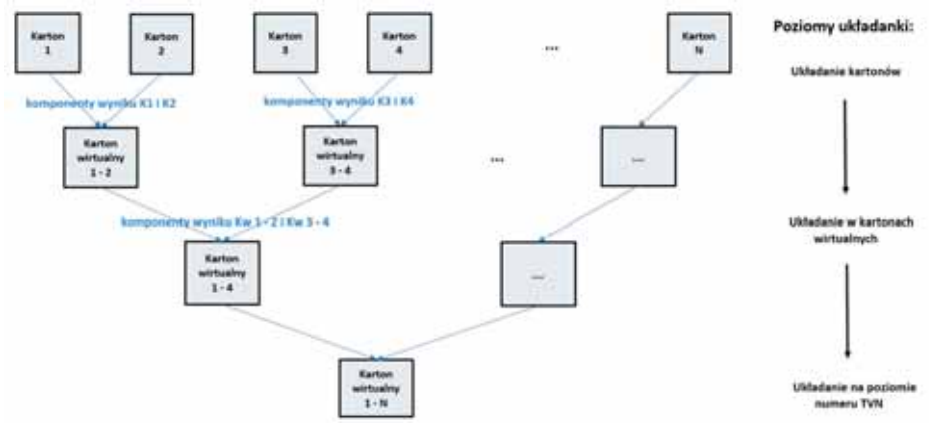
Ilustracja nr 8 prezentuje wycinek RECO-workflow. Przy preselekcji zestawiane są ilości i dopasowania skrawków i częściowych rekonstrukcji, które w oparciu o cechy ogólne i geometryczne są obiecującymi „kandydatami" do ułożenia. Łącznik konturów odpowiada za zgodność przebiegu krawędzi rozdarć wszystkich par fragmentów utworzonych w preselekcji. Części z niepasującymi konturami zostają odrzucone, wszystkie pozostałe są poddawane dalszej kontroli pod kątem „docelowego kształtu" i „łącznika kontekstu".

W dalszej części artykułu opisano przedstawione na ilustracji nr 8 etapy procesu RECO-workflow.

2.2.1. Model poziomy dla opracowania numeru TVN

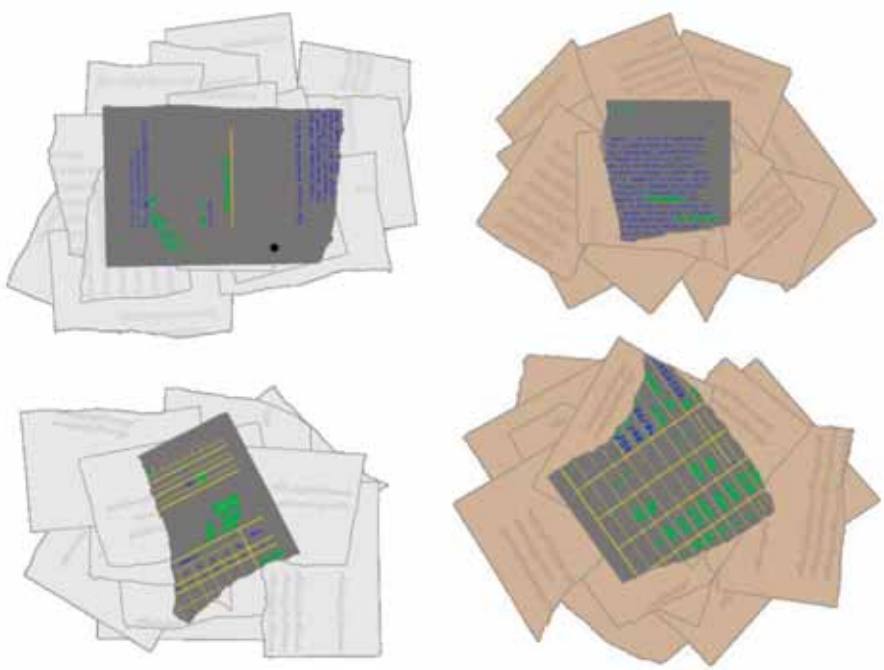
W trakcie wstępnego sortowania pracownicy Urzędu Pełnomocnika Federalnego ds. Materiałów Stasi z jednostek opracowania (TVe) wyjmują po kolei każdą warstwę skrawków i przenoszą je do kartonów w celu digitalizacji i archiwizacji. Przedstawiony na ilustracji nr 9 model poziomy dowodzi, że z większym prawdopodobieństwem będą do siebie pasować skrawki leżące obok siebie w jednostkach opracowania, niż fragmenty znajdujące się w różnych miejscach danej jednostki.

W związku z zasadą drzewa binarnego modelu poziomego wszystkie skrawki z poszczególnych kartonów są układane w pierwszej kolejności („układanie kartonów"). Utworzone częściowe rekonstrukcje oraz pozostałe nieułożone skrawki (czyli komponenty wyniku) zostają zebrane w tzw. wirtualnych kartonach, przy czym każdy z nich jest wypełniany komponentami wyników z dwóch sąsiednich kartonów. Następnie układane są elementy znajdujące się w tych wirtualnych kartonach (obrazowo: „jeden poziom wyżej", na ilustracji nr 9 wzdłuż czarnej strzałki). Powstałe w ten sposób komponenty wyniku zostają ponownie zebrane w wirtualnych kartonach. Zgodnie z tą zasadą zeskanowane skrawki grupuje się „na coraz to wyższym poziomie", aż wszystkie pozostałe komponenty układanki znajdą się w jednym kartonie („karton wirtualny 1-N" na ilustracji nr 9).



Ilustracja nr 9. Model poziomy dla opracowania numeru TVN (prezentacja uproszczona)

Źródło: Fraunhofer IPK, 2014



Ilustracja nr 10. Zawężenie przeszukiwanego obszaru na podstawie cech koloru, tekstury i kontekstu

Źródło: Fraunhofer IPK, 2014

2.2.2. Zawężenie przeszukiwanego obszaru/Zbiory przeszukiwanych obszarów (SRM)

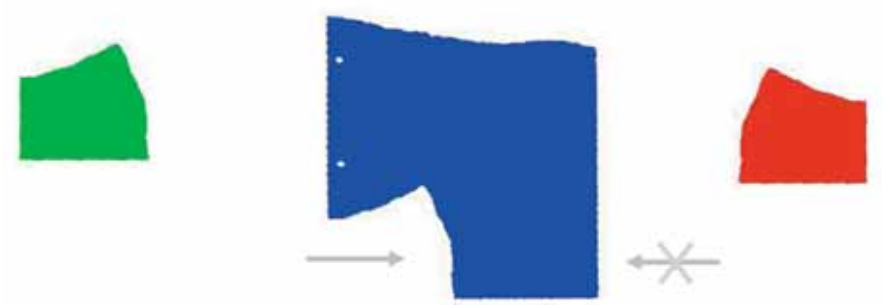
Dzięki odpowiedniej redukcji przeszukiwanego obszaru skrawki mające podobne właściwości obrazu zostają automatycznie zebrane w podzbiory, w tzw. zbiory przeszukiwanych obszarów. Na owe właściwości obrazu składają się np. kolor papieru lub rodzaj pisma na wszystkich komponentach układanki z jednego opracowywanego zbioru. W wyniku zawężenia przeszukiwanego obszaru mogłyby np. powstać zbiory z białymi, żółtymi, zielonymi i brązowymi elementami. Takie zbiory mogłyby zostać podzielone – zależnie od tego, jakie skrawki zawierają – jeszcze bardziej, chociażby na „białe pismo", „brak białego pisma", „żółte pismo", „brak żółtego pisma" itp.

Skrawki np. w jednym kolorze z większym prawdopodobieństwem będą do siebie pasować niż różnobarwne. Ponieważ tej ostatniej możliwości oczywiście nie można wykluczyć, sporządzanie zbiorów jest tylko pierwszym etapem opracowania, w którym liczy się na wysoką skuteczność w wyniku zestawiania prawdopodobnie pasujących fragmentów. W „ostatnim wirtualnym kartonie" w modelu poziomym kryteria redukcji obszaru poszukiwań zostają stopniowo złagodzone, aby w odpowiednich fazach procesu można było we wzajemnych porównaniach uwzględnić także różnokolorowe skrawki.

Zawężenie przeszukiwanego obszaru jest uruchamiane przed każdym opracowaniem wszystkich zbiorów. To znaczy, że odbywa się przed opracowaniem zarówno skrawków danego kartonu, jak i wirtualnych kartonów. Tym samym zbiór (SRM) składa się początkowo z najmniejszej liczby skrawków/częściowych rekonstrukcji, które są układane.

2.2.3. Preselekcja

Zadaniem preselekcji jest sprawdzenie, czy dwa skrawki i/lub dwie częściowe rekonstrukcje pasują do siebie geometrycznie, czy nie. Tym samym preselekcja służy w trakcie przebiegu pracy RECO-workflow jako filtr (filtr RECO), który automatycznie blokuje nieuzasadnione próby łączenia skrawków w trybie układania Puzzle-workflow (por. ilustracja nr 11). Zasadniczo preselekcję skonstruowano jako względnie tolerancyjny filtr.



Ilustracja nr 11. Preselekcja, odrzucenie niepasujących geometrycznie elementów układanki. Czerwony skrawek nie pasuje pod względem geometrii do niebieskiego fragmentu rekonstrukcji częściowej, niezależnie od tego, jak się go obraca lub przesuwa; natomiast zielony skrawek pasuje

Źródło: Fraunhofer IPK, 2014

2.2.4. Filtr RECO

W każdej dotychczasowej próbie dopasowania skrawków i/lub częściowych rekonstrukcji zastosowano następujące etapy procesu:

1. Dopasowywanie konturów.

Łącznik konturów weryfikuje, na ile pasują do siebie krawędzie dwóch komponentów. Wszystkie pary, które pasują do siebie, zostają poddane w następnych etapach dokładniejszym badaniom. Pary skrawków, których kontury nie przystają do siebie zbyt dobrze, zostają odrzucone i chwilowo nie są uwzględniane w trybie RECO-workflow.

Tym samym także łącznik konturów można postrzegać jako filtr RECO. W trybie RECO-workflow sprawdzane są bowiem wyłącznie fragmenty, których kontury dobrze pasują. Wszystkie pozostałe są natomiast odsiewane przez filtr RECO.

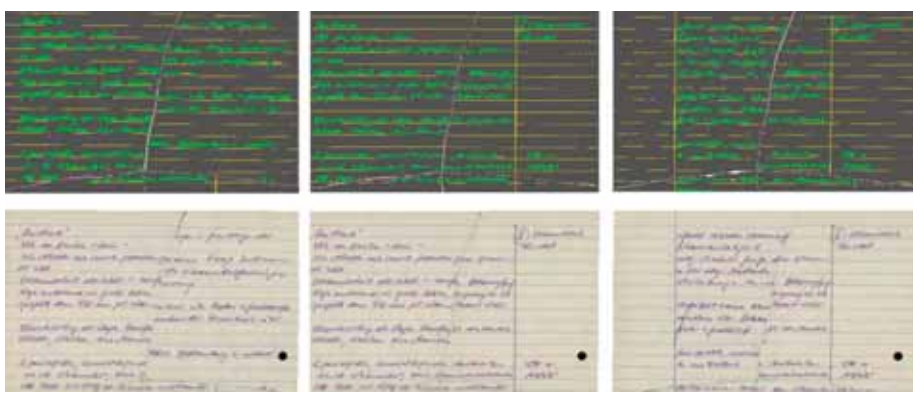
2. Kontrola kształtu docelowego.

W filtrze RECO-kontroler kształtu docelowego weryfikowane są wszystkie dopasowania, które przeszły przez filtr łącznika konturów. Badane jest, czy powstały wcześniej kształt jest dopuszczalny, czy też nie. Części o niemożliwych do zaakceptowania formach zostają chwilowo odrzucone, natomiast pozostałe poddaje się dalszym badaniom.

3. Dopasowanie kontekstu.

W filtrze RECO-łącznik kontekstu weryfikowane są wszystkie dopasowania, które przeszły przez filtr łącznika konturów oraz kontrolera kształtu docelowego. Sprawdzane jest, na ile obrazowe treści pasują do poszczególnych scaleń. Łącznik kontekstu weryfikuje także, czy treści obrazowe jednego komponentu układanki, jak wiersze, linie, kolory itp., kontynuowane są w następnym.

Wszystkie dopasowania dobrze dobrane pod względem treści obrazowych uznaje się za „kandydatów do celnych trafień" i poddaje ostatniemu badaniu – klasyfikacji strefowej – przeprowadzanemu w kolejnym etapie. Pozostałe elementy uznaje się za odrzucone i na razie nie poddaje filtrowaniu w trybie RECO-workflow.



Ilustracja nr 12. Interakcja pomiędzy dopasowaniem konturów i kontekstu. Dopasowanie konturów wszystkich trzech „kandydatów" (po 4 skrawki). Łącznik kontekstu zapobiega jednak scaleniu fałszywych „kandydatów" przedstawionych z lewej i prawej strony

Źródło: Fraunhofer IPK, 2014

2.2.5. Klasyfikacja strefowa

„Automatyczne scalenie dopasowania czy propozycja dla operatora?"

We wcześniejszych etapach procesu filtr RECO automatycznie określił wszystkie dopasowania jako „pasujące" lub „niepasujące", przy czym decyzja o poprawnym scaleniu została podjęta każdorazowo wyłącznie z poziomu danego filtra. W klasyfikacji strefowej jest określane, na ile dane dopasowanie jest odpowiednie dla wszystkich cech sprawdzonych w oparciu o poprzednie filtry.

Fragmenty określone przez klasyfikację strefową jako „łącznie dobrze pasujące" mogą zostać następnie scalone automatycznie, natomiast elementy ocenione jako „łącznie pasujące tylko częściowo" muszą przed ewentualnym scaleniem zostać zweryfikowane przez operatora.

Częściowa zgodność może wynikać stąd, że wprawdzie zarówno geometria, jak i kontury fragmentów znakomicie do siebie pasują, jednak obrazowa treść wykazuje niedopasowanie, np. nie ma kontynuacji między komponentami wierszy tekstu (por. ilustracja nr 12, z lewej i prawej strony). W tym przypadku łącznik kontekstu wprawdzie nie odrzuciłby całkowicie takiego zestawienia, ale i nie uznałby je za idealne, tak że klasyfikacja strefowa wydałaby łączną ocenę: „pasujące tylko częściowo".

Podsumowując, istnieją trzy grupy, do których przyporządkowany zostaje wynik dopasowania:

– grupa 1 – fragmenty pasują i mogą zostać scalone automatycznie;

– grupa 2 – fragmenty pasują tylko częściowo i przed scaleniem musi zostać poinformowany operator;

– grupa 3 – fragmenty nie pasują i dlatego zostają odrzucone.

2.2.6. Interaktywne stanowiska pracy z komputerem

Wszystkie propozycje drugiej grupy, a więc pasujące jedynie częściowo, są wyświetlane na stanowiskach pracy operatorów w tzw. podglądzie HQ (jest to element oprogramowania ePuzzlera). Do ich najważniejszych zadań należy zaakceptowanie lub odrzucenie propozycji scalenia drugiej grupy. Te same stanowiska pracy służą kontroli jakości, przy której wszystkie rekonstrukcje stworzone w jednej jednostce opracowania są sprawdzane pod względem poprawności oraz ewentualnie korygowane. Poniższe ilustracje przedstawiają przykłady „typowych sytuacji w podglądzie HQ".

Ilustracja nr 13. Interaktywne stanowiska pracy z komputerem – sprawdzanie wątpliwych dopasowań

Źródło: Fraunhofer IPK, 2014

Fragmenty przedstawione na ilustracji nr 13 (pięciofragmentowa rekonstrukcja na górze w zestawieniu z trzyfragmentową rekonstrukcją na dole; z prawej – łączenie dwóch skrawków) nie mogą zostać scalone automatycznie, ponieważ łącznik kontekstu nie ma wystarczających informacji (cech), aby podjąć bezbłędną decyzję. Na ilustracji zaprezentowanej z lewej strony całkowicie brakuje cech wzdłuż krawędzi rozdarcia zaznaczonej zieloną linią (puste skrawki). Na ilustracji przedstawionej z prawej strony dostępne są ogólne cechy (rodzaj pisma i kolor, *layout* akapitów itp.), jednak wzdłuż krawędzi rozdarć oznaczonej zieloną linią brak jest dla łącznika kontekstu „dowodu", że oba skrawki faktycznie pochodziły z tej samej strony. Może chodzić tutaj także o kilka podobnych do siebie formularzy podartych w całości, tak że dolny skrawek mógłby np. zostać przyporządkowany innej kartce. Dlatego ePuzzler kwalifikuje tego typu łączenia jako propozycję grupy drugiej, a następnie informuje o nich operatora.

Ilustracja nr 14. Interaktywne stanowiska pracy z komputerem – zapewnienie jakości

Źródło: Fraunhofer IPK, 2014

Ilustracja nr 14 przedstawia przykładowe rekonstrukcje wykonane (przynajmniej częściowo) automatycznie przez ePuzzlera. Muszą one zostać jeszcze skontrolowane przez operatorów i ewentualnie skorygowane, np. przez „oderwanie" pojedynczych skrawków i/lub częściowych rekonstrukcji od przesłanych prób połączeń. Ponadto, aby zrekonstruowane karty były bardziej czytelne, można dokonać dokładnych korekt pojedynczych elementów lub grup skrawków.

Ilustracja nr 15. Interaktywne stanowiska pracy z komputerem – zmiana klasyfikacji z częściowej rekonstrukcji na pełną

Źródło: Fraunhofer IPK, 2014

Ponieważ zgodnie z zasadą częściowe rekonstrukcje i nieużyte skrawki nie są przekazywane Urzędowi Pełnomocnika Federalnego (i tym samym nie będzie miał on dostępu do ich treści), wszystkie nieukończone odtworzenia w ramach jednego zbioru opracowania muszą zostać sprawdzone pod kątem kompletności treści. Jeżeli w częściowych rekonstrukcjach nie ma jedynie drobnych fragmentów lub brakujące części najprawdopodobniej nie zawierają żadnych treści, ich status może zostać podniesiony przez operatorów do rekonstrukcji pełnej. Ilustracja nr 15 przedstawia z prawej strony dwie tego typu nieukończone rekonstrukcje. We fragmencie zamieszczonym z lewej strony (beżowo-brązowy) decyzja o podniesieniu klasyfikacji jest relatywnie łatwa, ponieważ brakujący skrawek jest najprawdopodobniej pusty i tym samym nieistotny z punktu widzenia treści. Jeżeli chodzi o częściową rekonstrukcję znajdującą się z prawej strony (białej), decyzja nie jest tak jednoznaczna. Tutaj w brakującym miejscu pośrodku rekonstrukcji może znajdować się podpis, a więc jej klasyfikacja powinna zostać podniesiona – jeżeli w ogóle – dopiero po gruntownej archiwalnej ocenie. W czasie podejmowania tej decyzji należy zawsze pamiętać, że w związku z okolicznościami, w jakich doszło do przekazania (por. punkt 7: Układanka nie jest kompletna), nie jest możliwe do przewidzenia, czy brakujące części w ogóle istnieją. Nie wiadomo tym samym, czy pasujące fragmenty będzie można odnaleźć i dopasować. Jeżeli częściowa rekonstrukcja już zawiera istotne miejsca z tekstem, należy ją prawdopodobnie przenieść do analizy treści. W takim przypadku specjalista Urzędu Pełnomocnika Federalnego musi dokonać odpowiedniego „cięcia" i usunąć częściową rekonstrukcję z bieżącego procesu składania poprzez podniesienie klasyfikacji.

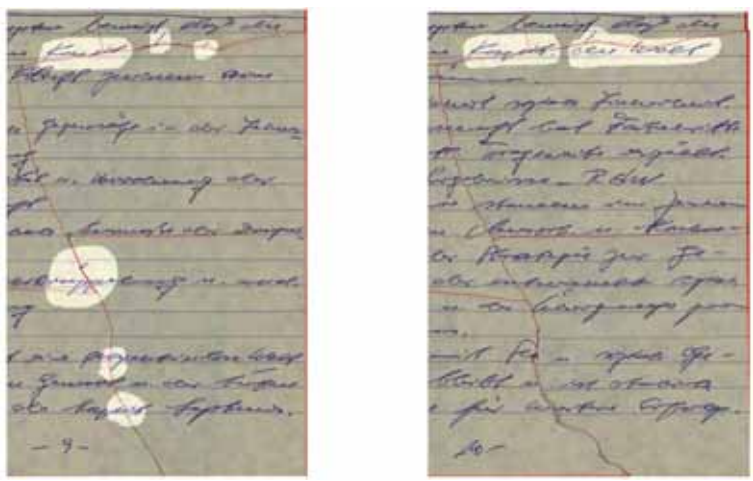
2.2.7. Granice automatycznej rekonstrukcji

Filtry RECO zaimplementowane w systemie ePuzzler osiągają granice swoich możliwości przede wszystkim w przypadku rękopisów, które zostały podarte w całości.

Ilustracja nr 16. Granice automatycznej rekonstrukcji. Dopasowanie konturów wszystkich trzech „kandydatów" (po 8 skrawków). Ponieważ łącznik kontekstu nie dysponuje jednoznacznymi cechami, żaden z „kandydatów" nie zostałby odrzucony automatycznie

Źródło: Fraunhofer IPK, 2014

Na ilustracji nr 16 można dostrzec dwie fałszywe i jedną poprawną propozycję łączenia fragmentów. Na pierwszy rzut oka wszystkie projekty wydają się poprawne, ponieważ na brzegach nie ma jakichkolwiek istotnych złamań kontekstu. Dopiero zrozumienie tekstu identyfikuje obie lewe propozycje jako fałszywe, a prawą – jako właściwą. Na niepoprawne dopasowanie elementów wskazuje błędna numeracja stron (u góry i na dole) na obu kartkach po lewej stronie. Częściowe powiększenie fragmentu z ilustracji nr 16, znajdującego się pośrodku i z prawej strony, pokazuje dalsze – także wymagające zrozumienia treści – cechy kontekstu (por. ilustracja nr 17). Wynika z nich, że kontynuacja tekstu wzdłuż krawędzi rozdarcia środkowego jest niepoprawna, natomiast wzdłuż krawędzi rozdarcia prawego – właściwa.



Ilustracja nr 17. Granice automatycznej rekonstrukcji. Identyfikowanie fałszywej (z lewej) i poprawnej (z prawej) kontynuacji kontekstu wzdłuż krawędzi rozdarcia wymaga zrozumienia tekstu

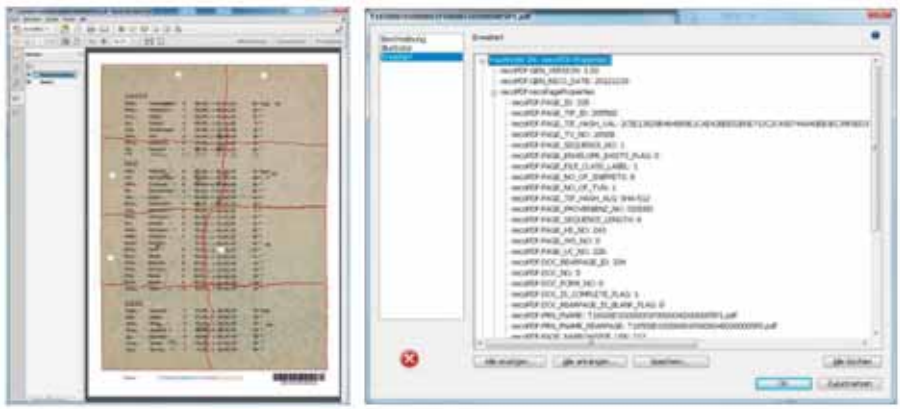
Źródło: Fraunhofer IPK, 2014

W związku z brakiem algorytmiki, która sprawia, że dowolne teksty stają się czytelne dla komputera, do systemu ePuzzler nie zaimplementowano żadnej aplikacji służącej rozpoznawaniu tekstu. Wszystkie bowiem fragmenty przedstawione na powyższych ilustracjach zostałyby automatycznie scalone. Jeżeli „kandydat" umiejscowiony z prawej strony na ilustracji nr 16 zostałby (przypadkowo) opracowany jako pierwszy w trybie RECO-workflow, powstała rekonstrukcja byłaby poprawna. W innych kombinacjach utworzone odtworzenia okazałyby się niewłaściwe.

2.3. Przygotowanie wyników rekonstrukcji

Wyniki odtworzeń przekazywane są Urzędowi Pełnomocnika Federalnego ds. Materiałów Stasi w formie wielowarstwowych plików Multi-Layer-PDF/A na taśmach magnetycznych LTO. Obrazy rekonstrukcji zapisywane są w dokumentach PDF/A jako pliki

JPEG2000. Opcjonalnie na obrazach rekonstrukcji można uwidocznić krawędzie rozdarć i numery skrawków. Poza obrazami pliki PDF/A zawierają obszerne rekordy metadanych w formacie XMP, z których można skorzystać przy cyfrowym formatowaniu odtworzonych stron[5]. Każda zrekonstruowana karta ma wygenerowane dwa pliki PDF/A: jeden dla awersu, drugi dla rewersu. Poza obrazem rekonstrukcji (ewentualnie nieznacznie zmniejszonym) każdy plik PDF/A zawiera z prawej dolnej strony kod z oznaczeniem przypisanym dla każdego obrazu (por. ilustracja nr 18).



Ilustracja nr 18. Multi-Layer-PDF/A i przynależne metadane

Źródło: Fraunhofer IPK, 2014

Wszystkie pliki PDF/A z rekonstrukcją automatycznie oznaczoną przez ePuzzler jako zapełnioną są drukowane (por. ilustracja nr 19). Jeżeli rewers rekonstrukcji jest pusty, jego plik PDF/A nie zostaje wydrukowany, a PDF/A awersu odpowiednio się oznacza.

Ilustracja nr 19. Wydruk „małej" jednostki TVN (niemalże 3600 stron wydruku)

Źródło: Fraunhofer IPK, 2014

[5] Formatowanie komputerowe nie jest elementem funkcjonalności systemu ePuzzler.

Analogicznie do PDF/A każda wydrukowana strona – poza obrazem rekonstrukcji (ewentualnie nieznacznie zmniejszonym) – zawiera na dole z prawej strony kod paskowy z oznaczeniem unikatowym dla każdego wydruku. Dzięki temu oznaczeniu każda wydrukowana strona może zostać przyporządkowana plikowi w systemie informatycznym Urzędu Pełnomocnika Federalnego, tak aby można było m.in. ustanowić jednoznaczne połączenie pomiędzy (analogicznym) wydrukiem a (elektronicznymi) metadanymi danej rekonstrukcji.

3. Pozostałe scenariusze zastosowania

Zautomatyzowana wirtualna rekonstrukcja jest wszechstronnym narzędziem, którego możliwości zastosowania wykraczają poza odtworzenie akt Stasi. Może być ono np. stosowane do efektywnego zachowywania i restauracji dokumentów i obiektów ważnych zarówno w aspekcie kulturowym, jak i ogólnospołecznym. Istotny element stanowią tutaj systemy wymagające asysty operatora, wykorzystujące wiedzę ekspertów w procesie odtwarzania i umożliwiające ponadto fizyczną rekonstrukcję dóbr kultury.

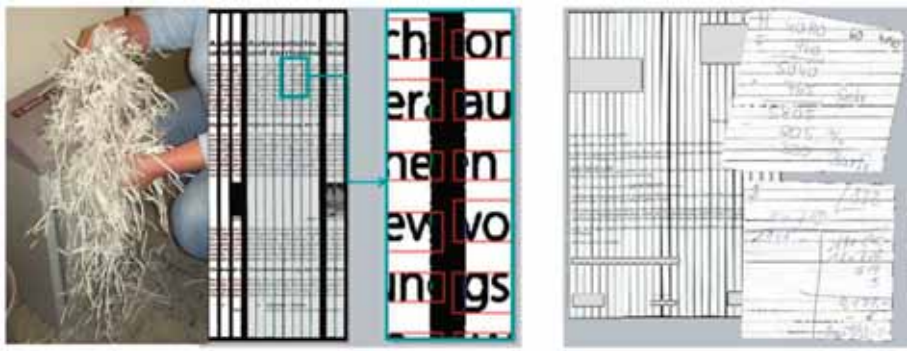
W przypadku systemów rekonstrukcji bazujących na asyście rozróżnia się dwa podejścia. Pierwszym z nich jest koncepcja wirtualnej rekonstrukcji bazującej na asyście. Celem – podobnie jak w zautomatyzowanej wirtualnej rekonstrukcji – jest cyfrowe odtworzenie treści. Fizyczne fragmenty po przeprowadzeniu digitalizacji nie są z reguły potrzebne i zostają zarchiwizowane. System ten jest ukierunkowany na fragmenty, których całkowite automatyczne odtworzenie nie jest technicznie możliwe i dlatego też rekonstrukcja wymaga współpracy ze strony człowieka. Jako przykład służy odtwarzanie pociętych w niszczarce dokumentów lub odczytywanie zniszczonych samochodowych tablic rejestracyjnych.

Systemy wspierające fizyczne rekonstrukcje znajdują się obecnie w opracowaniu. Tutaj wynik wirtualnego odtworzenia ma służyć jako podstawa do dalszej manualnej rekonstrukcji lub restauracji. Systemy tego typu stawiają wysokie wymagania przebiegowi pracy zarówno przed digitalizacją, jak i po wirtualnej rekonstrukcji. Ponieważ celem tego założenia jest fizyczne odtworzenie uprzednio zdigitalizowanych obiektów, konieczne jest posiadanie wydajnego i przystosowanego do zadania systemu śledzenia. Niezależnie od założonego celu w wielu przypadkach fizycznej rekonstrukcji nieodzowna jest analiza ekspertów. Z jednej strony może chodzić o specjalistyczne umiejętności archiwisty lub konserwatora, np. wiedza na temat miejsca znaleziska, wieku lub ówczesnych okoliczności, która *a priori* jest wprowadzana do wirtualnej rekonstrukcji jako metainformacje. Z drugiej zaś strony może chodzić o wnioski i spostrzeżenia, które pojawiają się dopiero w trakcie procesu cyfrowej rekonstrukcji. Mogą one dotyczyć np. niewidocznych na małych elementach treści, które są dostrzegalne dopiero na częściowych rekonstrukcjach złożonych z kilku fragmentów. Komputer wciąż nie jest w stanie dorównać wiedzy eksperta bazującej na wieloletnim doświadczeniu, jest ona zatem nieodzowna do uzyskania wysokich współczynników rekonstrukcji w procesie odtwarzania.

Założenie to wpływa obecnie na rozwój dwóch systemów wspierających fizyczną rekonstrukcję, które zostały opisane w części artykułu „Systemy wspierające fizyczną rekonstrukcję".

3.1. Wspierana wirtualna rekonstrukcja dokumentów pociętych w niszczarce

Odtworzenie dokumentów pociętych w niszczarce stanowi przypadek szczególny. Proces wyselekcjonowania cech i dopasowania poddany jest tutaj drobiazgowym wymogom – jednolitość konturów fragmentów, tak więc rekonstrukcja może opierać się wyłącznie na cechach treści i przebiegu kolorów. Ponadto fragmenty mają szerokość i długość wynoszącą jedynie kilka milimetrów, zatem cechy odnoszące się do treści można wyodrębnić tylko na podstawie zaledwie kilku pikseli (por. ilustracja nr 20).



Ilustracja nr 20. Rekonstrukcja dokumentów pociętych w niszczarce. Od lewej do prawej: przygotowanie materiałów dowodowych, dopasowanie kontekstu, wirtualna rekonstrukcja

Źródło: Fraunhofer IPK, 2014

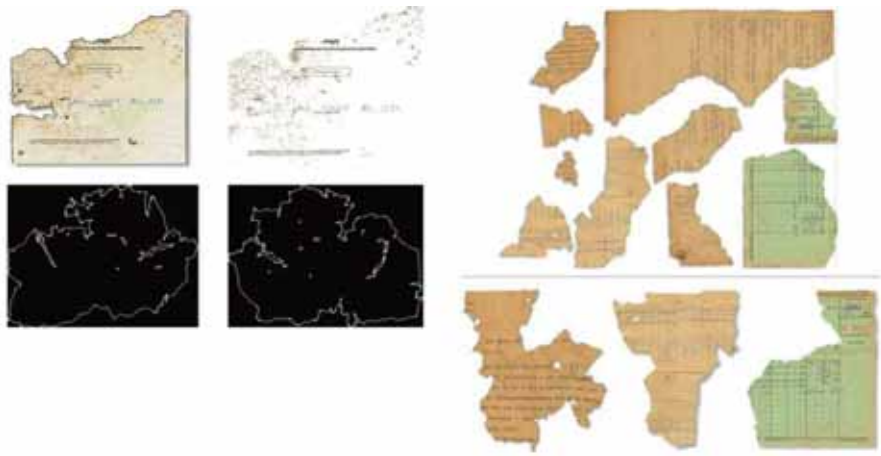
W czasie rekonstrukcji dokumentów pociętych w niszczarce rezygnuje się z przyporządkowywania treści do klas takich jak „pismo" czy „liniowanie". Zamiast tego wszystkie elementy na jednym paśmie pociętym w niszczarce traktowane są – łącznie z kolorem papieru – jako obiekty geometryczne, których rozproszenie jest ustalane matematycznie według reguł binarnego dopasowania *Stringmatching* i prawdopodobieństwa. Stąd wynik jest geometrycznie możliwy, ale niekoniecznie właściwy, jeżeli chodzi o treść. Dlatego rekonstrukcja dokumentów z niszczarki wymaga stałej kontroli człowieka.

3.2. Systemy wspierające fizyczną rekonstrukcję

3.2.1. Rekonstrukcja zniszczonych zbiorów archiwalnych z Miejskiego Archiwum Historycznego w Kolonii

Gmach Miejskiego Archiwum Historycznego w Kolonii oraz dwa sąsiednie budynki mieszkalne runęły 3 marca 2009 r. Około 90 proc. materiałów archiwalnych zostało zasypanych, z czego do dzisiaj udało się uratować ponad 85 proc. Od tego czasu Miejskie Archiwum Historyczne w Kolonii stoi przed wyzwaniem polegającym na oczyszczeniu i odrestaurowaniu milionów fragmentów, a także przywróceniu dokumentów archiwalnych do stanu pierwotnego. Liczba oraz stopień uszkodzeń materiałów, a także brak kry-

teriów ograniczających poszukiwania nie pozwalają w tym przypadku na czysto manualną rekonstrukcję. System wspierający oparty na technice ePuzzlera mógłby znacznie przyczynić się do odtworzenia i zabezpieczenia dużych części materiałów archiwalnych, które są ważne ze względów historycznych, lecz zostały poważnie uszkodzone w wyniku działania sił mechanicznych w trakcie zawalenia się budynków.



Ilustracja nr 21. Zanieczyszczone i uszkodzone fragmenty Miejskiego Archiwum Historycznego w Kolonii. Z lewej: zanieczyszczony fragment (u góry), analiza konturów (na dole). Z prawej: fragmenty (u góry), wirtualna rekonstrukcja (na dole)

Źródło: Fraunhofer IPK, 2014

Możliwość stosowania procesu wirtualnej rekonstrukcji została dowiedziona na zlecenie Kolonii już na początku 2010 r. Wykorzystano wtedy reprezentatywną próbę losową obejmującą 1000 elementów. Obecnie Instytut Fraunhofera wraz z firmą MusterFabrik z Berlina, na zlecenie Kolonii i przy specjalistycznej współpracy z Miejskim Archiwum Historycznym, opracowuje prototypowe rozwiązanie do wirtualnej rekonstrukcji wybranych materiałów archiwalnych. Jednocześnie powstaje koncepcja dotycząca oczyszczania, digitalizacji i rekonstrukcji wszystkich fragmentów, które udało się uratować w Kolonii. Projekt ten jest finansowany ze środków Unii Europejskiej (Europejski Fundusz Rozwoju Regionalnego).

3.2.2. Odtworzenie fragmentów szklanej mozaiki w kaplicy

Pola, na których można stosować zautomatyzowaną wirtualną rekonstrukcję, nie są ograniczone wyłącznie do materiałów dwuwymiarowych, np. jak papier. W przeszłości były prowadzone już prace w Instytucie Fraunhofera nad koncepcją zautomatyzowanej wirtualnej rekonstrukcji trójwymiarowych obiektów z płaską powierzchnią. Nadają się one do pierwszych prób w tym zakresie, ponieważ kompleksowy trójwymiarowy zapis fragmentów nie jest konieczny. Zamiast tego dąży się do rekonstrukcji za pomocą infor-

macji dotyczących powierzchni. Dane uzyskuje się dzięki analizie konturów i tekstury analogicznie do zakresów 2D. Dodatkowo – i to wykracza już poza podejście charakterystyczne dla rekonstrukcji 2D – uwzględnia się grubość krawędzi złamań, aby wspomóc poszukiwanie zgodności poszczególnych fragmentów. Ten sposób postępowania jest pierwszym krokiem do „prawdziwej" rekonstrukcji 3D: analizowane jest zdjęcie 2D jednej strony obiektu i uzupełniane o informacje dotyczące głębi. Dlatego też można ten proces interpretować jako rekonstrukcję 2,5D. Z wielu zdjęć 2,5D, które odwzorowują wszystkie obszary powierzchni fragmentów, można w kolejnym kroku utworzyć pełne obiekty 3D. Największe wyzwanie polega na tym, aby właściwie odseparować od siebie powierzchnie zarówno obiektów, jak i złamań, a następnie wygenerować ich odpowiednie widoki.

Warunkiem dla każdej formy przestrzennej wirtualnej rekonstrukcji są urządzenia rejestrujące z wystarczająco dobrą jakością odzwierciedlenia, które nadają się do zastosowania do poszczególnych materiałów przeznaczonych do rekonstrukcji. Ponadto etap procesu polegający na rejestracji obrazu dodatkowo wydłuża czas opracowywania obiektów trójwymiarowych. Z tego powodu konieczne jest każdorazowe opracowywanie koncepcji dostosowanej do danego zadania, która taki etap na tyle zautomatyzuje i skutecznie przeprowadzi, na ile jest to oczywiście możliwe.

Obecnie dzieje się tak przy projekcie cyfrowej rekonstrukcji dóbr kultury w 2,5D, nad którym pracują Instytut Fraunhofera i firma MusterFabrik. Celem przedsięwzięcia jest techniczne badanie i prototypowa realizacja algorytmów rekonstrukcyjnych 2,5D oraz niezbędnych urządzeń peryferyjnych (skanery 2,5D, Viewer Appliance itp.) przeznaczonych do rekonstrukcji i repozycji fragmentów dóbr kultury w 2,5D. Za materiały referencyjne dla pierwszych testów prototypowych narzędzi służą fragmenty mozaiki szklanej z byłej kaplicy pochówkowej w miejscu Buchholz w Bredereiche, Fürstenberg nad Hawelą.

Ilustracja nr 22. Fizyczna rekonstrukcja fragmentów mozaiki szklanej. Z lewej: uszkodzony fresk ścienny w kaplicy; z prawej: odnalezione fragmenty

Źródło: Fraunhofer IPK, 2014

Z lewej strony na ilustracji nr 22 nad trzema oknami jest widoczna część zachowanego fresku ściennego. Pierwotnie rozpościerał się on pomiędzy oknami oraz okalał je na zewnątrz, niemalże do podłogi kaplicy. Z prawej strony można dostrzec niewielką część

odnalezionej mozaiki szklanej. Poza dość dużymi fragmentami jest też wiele małych kawałków (por. z monetą 1 euro), które sprawiają, że manualna rekonstrukcja – o ile w ogóle jest możliwa – okazuje się wyjątkowo czasochłonna.

Jednym z celów projektu jest w pierwszej kolejności opracowanie techniki digitalizacji 2,5D nadającej się do masowego wykorzystania, a także budowa prototypu. Poza zapisywaniem informacji o głębiach wyzwaniem jest elektroniczna rejestracja szklanych powierzchni mozaiki z zachowaniem wiernego odzwierciedlenia kolorystki i geometrii. W kolejnym etapie ma zostać zaimplementowany prototypowy system wspierający, który będzie proponował restauratorowi pasujące do siebie fragmenty mozaiki. Dzięki wsparciu inteligentnego systemu śledzenia, który jest opracowywany równolegle do właściwej digitalizacji, restaurator może zweryfikować za pomocą „prawdziwych" (fizycznych) kawałków, które z propozycji systemu wspierającego są poprawne, a następnie je skompletować, choć obecnie tylko wirtualnie. Jeżeli zachowała się wystarczająca liczba pasujących elementów, wirtualny fresk może być sukcesywnie odwzorowywany. Ponieważ nie będzie miał przy oknach odpowiednich dopasowań, system wspierający utworzy w ten sposób cały obraz rekonstrukcji, który umożliwi jednoznaczne i wierne oryginałowi pozycjonowanie kawałków na ścianie. Restaurator może także pobrać z systemu wspierającego odpowiedni „plan budowy", ponieważ sposób połączenia poszczególnych kawałków z sąsiednimi jest zapisany w formie elektronicznej.

Streszczenie

Pełnomocnik Federalny do spraw Materiałów Państwowej Służby Bezpieczeństwa NRD przejął dużo zniszczonych materiałów archiwalnych Stasi (podarte, pocięte w niszczarce) i szukał rozwiązań, które umożliwiają szybszą rekonstrukcję ww. dokumentów. W 2007 r. podjął współpracę z Instytutem Fraunhofera do spraw Systemów Produkcyjnych i Technik Konstrukcyjnych w Berlinie, który opracował kompleksowy system umożliwiający zautomatyzowaną wirtualną rekonstrukcję zniszczonych dokumentów. Proces obejmuje digitalizację, układanie oraz opracowanie wyników. Program „ePuzzler" analizuje różne cechy skrawków (kontury, kolor papieru, pismo, liniowanie) i składa zeskanowane fragmenty dokumentów w strony, a następnie w jednostki archiwalne. Podczas digitalizacji tworzony jest obraz skrawka i przeniesione zostają wszystkie rozpoznane informacje przydatne do dalszego opracowania. Obrazy skrawków danej jednostki opracowania zostają wgrane do systemu ePuzzler, a następnie są układane według strategii drzewa binarnego. Skrawki mogą wykazywać duże zróżnicowanie w obrębie ich cech, dlatego wariancja liczby wszystkich skrawków jest niezwykle duża. Rekonstrukcje są kontrolowane pod względem poprawności i korygowane. Na koniec opracowania danej jednostki wszystkie rekonstrukcje są poddane manualnemu procesowi zapewnienia jakości. Po jego zakończeniu wszystkie rekonstrukcje są automatycznie konwertowane przez ePuzzlera do formatu archiwizacji PDF/A.

Słowa kluczowe: ePuzzler, digitalizacja, wirtualna rekonstrukcja dokumentów, cyfrowa rekonstrukcja dóbr kultury, odtwarzanie zniszczonych materiałów archiwalnych, rekonstrukcja dokumentów Stasi, Instytut Fraunhofera do spraw Systemów Produkcyjnych i Technik Konstrukcyjnych, Urząd Pełnomocnika Federalnego do spraw Materiałów Państwowej Służby Bezpieczeństwa NRD, MusterFabrik Berlin; Miejskie Archiwum Historyczne w Kolonii.

Summary

Federal Attorney for the issues of the Materials of the State Security Service of the former German Democratic Republic undertook numerous destroyed archival materials of Stasi (torn, shredded) and searched for solutions that allowed for faster reconstruction of the above-mentioned documents. In 2007 he started cooperation with the Fraunhofer Institute for Production Systems and Design Technology in Berlin that prepared the complex system allowing for automated virtual reconstruction of the destroyed documents. The process includes digitisation, arrangement and preparation of the results. 'ePuzzler' software analyzes various features of the fragments (contours, colour of the paper, handwriting, linearization) and arranges the scanned fragments of the documents into the pages and, subsequently, into the archival units. During digitisation the image of the fragments is created and all recognized information useful for further preparation is transferred. The images of the fragments of a specific unit of the preparation are uploaded to ePuzzler software and afterwards they are arranged according to the strategy of the binary tree. The fragments may display a significant diversification within their features, consequently the variance of the number of all fragments is extremely huge. Reconstructions are checked taking into account accuracy and corrected. At the end of preparation of a particular unit all reconstructions are subject to the manual process of quality assurance. After its completion all reconstructions are automatically converted to the PDF/A archivization format by ePuzzler.

Key words: ePuzzler, digitisation, virtual reconstruction of documents, digital reconstruction of cultural objects, restoration of destroyed archival materials, reconstruction of Stasi documents, the Fraunhofer Institute for Production Systems and Design Technology, the Office of Federal Attorney for the issues of the Materials of the State Security Service of the former German Democratic Republic, MusterFabrik Berlin; Historical Archive of the City in Cologne.