

**Институт по математика и информатика,
Българска академия на науките**

ИЗИСКВАНИЯ КЪМ ЦИФРОВИЗАЦИЯТА НА СРЕДНОВЕКОВНИ СЛАВЯНСКИ РЪКОПИСИ

разработени по Изследователски проект „Развитие на кирилицата от IX до XIV век в южнославянските земи: изследване и цифрово представяне“, Договор КП-06-Н90/9 от 11.12.2024 с Национален фонд „Научни изследвания“, Конкурс за финансиране на фундаментални научни изследвания – 2024 г.

Автори: проф. д-р Десислава Панева-Маринова, проф. д-р Радослав Павлов, доц. д-р Детелин Лучев, доц. д-р Максим Гойнов, доц. д-р Лилия Павлова (Институт по математика и информатика, Българска академия на науките), доц. д-р Янислав Желев (Институт по математика и информатика, Българска академия на науките; Бургаски свободен университет)

2026, София

СЪДЪРЖАНИЕ

1. Въведение. Анализ на състоянието и особености на избрани колекции със славянски ръкописи	4
2. Дигитализация на славянски ръкописи. Задачи и насоки на процеса.....	6
3. Анализ и препоръчителни методи, техники и технически средства за цифрово заснемане/сканиране с оглед осигуряване на цифрово съдържание, подходящо за анализ, както от експерти в областта, така и чрез средствата на изкуствения интелект и други видове машинен анализ	10
3.1. Подходи за дигитализация	12
3.1.1. Класификация на скенерите	13
3.1.2. Показатели за оценка на скенери	18
3.1.3. Интерфейс.....	20
4. Технически характеристики за високо качество на цифровото изображение (резолюция, дълбочина на пикселите, цветови модел и формат за съхранение на файлове и др.).....	22
4.1. Описание на форматите за съхранение на изображения.....	23
4.1.1. Векторни графични изображения и формати	23
4.1.2. Растерни графични изображения и формати	24
5. Препоръчителен софтуер и настройки. Работата с избран софтуер за цифрова обработка	25
5.1. Софтуер	26
5.2. Софтуерни изисквания – програмни средства за дигитализиране на изображения 27	
5.3. Дигитализация и обработка на изображения. Формати за съхранение на дигитализираните обекти.....	27
5.4. Описание на форматите за съхранение и представяне на изображения в уеб (JPEG, GIF, PNG, DJVU).....	28
5.4.1. Файлов формат JPEG.....	29
5.4.2. Файлов формат GIF	29
5.4.3. Файлов формат PNG	30
5.4.4. Файлов формат DJVU	30
5.5. Описание на форматите за съхранение на изображения в архив (PDF, BMP, TIFF) 31	
5.5.1. Файлов формат PDF	32
5.5.2. Файлов формат BMP	32
5.5.3. Файлов формат TIFF.....	33
6. Особенности на използваните файлови формати за цифрови колекции от славянски ръкописи и техники за цифровата обработка на изображенията	34

6.1.	TIFF спрямо JPEG файлове	34
6.2.	JPEG спрямо PDF файлове	36
7.	Архивиране и представяне на цифрови колекции от славянски ръкописи.....	38
7.1.	Обработка на дигитализираните документи	38
7.2.	Комплектуване на файловете.....	39
7.3.	Именуване и подреждане на файловете.....	39
8.	Настройване на техниката	42
9.	Подходящи стандарти и спецификации за описание на цифровите обекти; изграждане на цифрови колекции/цифрови хранилища	44
9.1.	Индексиране на цифровите обекти. Метаданни	44
9.1.1.	Стандарт Dublin Core	44
10.	Характеристики за цифрова обработка на изображенията според последващи използване и достъп.....	48
11.	Насоки в развитието на дигитализацията на славянски ръкописи.	51
12.	Техники на изкуствения интелект за разпознаване на ръкописен текст	54
13.	Заключение	57

1. Въведение. Анализ на състоянието и особености на избрани колекции със славянски ръкописи

Анализът на състоянието и особеностите на средновековни славянски ръкописи започва от признаването на тяхната висока историческа, културна и лингвистична стойност и необходимостта от съхранение и широк достъп чрез съвременни цифрови технологии. Цифровизацията не е единствено процес на сканиране - тя включва комплекс от дейности: консервация преди заснемане, висококачествена дигитална фотография (включително мултиспектрално изобразяване), метаданни и семантично описание, транскрипция, лингвистична и палеографска анотация, както и интеграция в дигитални хранилища и ресурси за изследователи и публика.

Сред специфичните предизвикателства са разнообразието от писмени системи и локални варианти на старославянския език, нестандартизирани орфографски и диакритични форми, ръкописни корекции и палимпсести, които затрудняват автоматизираните методи като OCR. Техническите и семантични стандарти (напр. TEI, PFI) са ключови за интероперабилност и дългосрочна устойчивост, но прилагането им изисква тесни взаимодействия между филолози, палеографи, консерватори и технологични специалисти. Важно е и внимателното управление на правата и репрезентацията при дигиталните копия, особено при уникални и чувствителни материали.

Цялостният анализ включва оценка на наличните колекции, качество и унификация на метаданните, наличието на емпирични инструменти за транскрипция и търсене, както и възможностите за образователно и научно използване. Заклученията и препоръките трябва да насочват към приоритизиране на материали за дигитализация, внедряване на общи стандарти, развитие на адаптирани инструменти на изкуствения интелект за старославянски текстове и укрепване на интердисциплинарните мрежи за поддръжка и устойчиво управление на дигиталните архиви.

Текущ анализ на състоянието и особености на избрани колекции със славянски ръкописи са представени в сп. Старобългаристика¹. Той показва, че по-голяма част от кодексите имат само кратко описание, придружено от микрофилмови или дигитални фотографии, което не дава изчерпателна

¹ Скарпа, М., Рипаранте, М., Панева-Маринова, Д., Гойнов, М., Лучев, Д. 2022. Онлайн база данни на южнославянски ръкописи от XIV в. Старобългаристика, XLVI (2022), 1, 181-202.

информация, а само библиографски препратки за описанието на ръкописите. Съществуващите изключения, като дигиталните каталози със славянски ръкописи на Руската национална библиотека, на Държавния исторически музей в Москва, на Националните библиотеки на Чехия, Полша и Словения и др. са опити в търсене на начини и възможности за представяне на наличните архиви.

Освен през изложения в посочената по-горе статия поглед към дигиталните каталози, те могат да бъдат анализирани от гледна точка на цифровото представяне на дигитализираните документи и използваните формати.

В дигиталния каталог на славянските ръкописи ² на Руската национална библиотека (РГБ) за представяне на документите са използвани файлове в PDF/A стандарт. Същевременно в уеб сайт ³, съдържащ дигитализирани ръкописи РГБ и Централния държавен архив (ЦДА) в Москва, ръкописите на РГБ са дадени по страници в JPG формат в два варианта с различна компресия (в съотношения до 1:20), като има и възможност за изтегляне на PDF файла на съответни ръкопис (с големина от порядъка на стотици мегабайти).

По подобен начин е решен достъпът до съдържанието на славянските ръкописи в каталога на Държавния исторически музей (ДИМ) в Москва⁴, като в случая за преглед на съдържанието се използват JPEG файлове с изключително висока компресия на някои от страниците или на всички страници на съответния ръкопис, като за част от ръкописите съществува възможност и да бъдат изтеглени в PDF формат в два варианта с различна компресия. Някои от документите са сканирани на плосък скенер, което е довело до изрязване на краищата на страници и загубата на съдържателна информация в дигиталното копие спрямо оригинала. За разлика от ДИМ в електронния каталог на ръкописите в Руската национална библиотека (РНБ)⁵ са представени само в JPEG файлове с малка компресия и само в някои случаи са сканирани не на двойни страници, а страница по страница (като Зографското евангелие, например).

² Каталог Записей Российской государственной библиотеки: <https://search.rsl.ru/#ef=1&af=1&yf=1300&yt=1399&c=dlruk&l=777>

³ Собрание рукописей и старопечатных книг: <https://lib-fond.ru/>

⁴ Государственный исторический музей: https://catalog.shm.ru/entity/OBJECT?fund_ier=647760263_647760288

⁵ Российская национальная библиотека: <http://nlr.ru/manuscripts>

В дигиталния каталог на Националната библиотека на Чехия⁶ се използват изключително компресирани JPEG файлове на страниците на дигитализираните ръкописи и книги, които изпълняват ролята на своеобразни икони за отваряне на големи PNG файлове. По подобен начин се извършва представянето на славянските ръкописи и Националната библиотека на Полша⁷, докато в Националната и университетска библиотека на Словения⁸ наблюдаваме използването на големи PDF файлове с висока резолюция.

В своята цялост проучването на използваните формати в бази данни за славянски ръкописи подсказва наличието на файлове на сканираните ръкописи с много висока резолюция и последващата им компресия и конвертирането във формати, съответстващи на целите на използването и онлайн достъпа до тях.

2. Дигитализация на славянски ръкописи. Задачи и насоки на процеса

В условията на ускорена цифрова трансформация дигитализацията се утвърждава като стратегически процес за опазване, съхраняване, управление и предоставяне на достъп до информационни, научни и културни ресурси. Тя играе ключова роля за създаването на съвременни цифрови екосистеми, които осигуряват ефективно управление на знанията, оперативна съвместимост между различни информационни системи и широк достъп до цифрово съдържание. Дигитализацията не се ограничава единствено до преобразуването на аналогови обекти в цифров формат, а представлява основа за изграждането на интелигентни услуги за търсене, анализ, семантично описание, автоматизирана обработка и повторна употреба на цифровите ресурси. Съвременните технологии, включително изкуственият интелект, обработката на естествен език, компютърното

⁶ Manuscriptorium, дигитален каталог на Националната библиотека на Чешката република: <http://www.manuscriptorium.com/en>

⁷ Дигитален каталог на Националната библиотека на Полша: https://polona.pl/search/?filters=metatype:r%C4%99kopisy_cyrylickie&sort=date%20asc

⁸ Digitalna knjižnica Slovenije: <http://dlib.si/results/?query=%27col%3dKopitarjeva+zbirka%27&desc=URN%3aNBN%3aSI%3acol-5AJX3A3O&pageSize=25&flanguage=chu>

зрение и големите езикови модели, значително разширяват възможностите за извличане на знания и създаване на нова добавена стойност от дигитализираните данни. В областта на културното наследство, науката и образованието дигитализацията създава условия за дългосрочно съхраняване на оригиналните обекти, улеснява тяхното изследване и популяризиране, подпомага развитието на цифровите хуманитарни науки и насърчава използването на цифровите ресурси в научни изследвания, образователни дейности и иновативни обществени приложения.

Дигитализиране или дигитализация представлява създаване на обект, изображение, звук, документ или сигнал от дискретен набор от неговите точки или проби. Строго погледнато, дигитализиране означава просто улавяне на аналогов сигнал в цифров вид, но дигитализиране на текстове се използва също и за означаване на процеса на конвертиране на изображение в текст чрез технология за разпознаване на символи. Резултатът се нарича "дигитално изображение" на обекта в "цифров вид" на сигнала.

Процесът на дигитализация позволява информация от всякакъв вид, във всякакъв формат да бъде поддържана непрекъснато във времето със същото качество и ефикасност. Процесът на дигитализация оставя впечатление на не толкова сложен, тъй като за него може да се използва не особено мощен компютър и потребителски скенер със стандартен софтуер за създаване на дигитално копие. В действителност обаче процесите са специфични, особено ако за крайния резултат е важно качеството на файла, формата, използван за съхранение и опазването му в бъдеще, неговото описание, идеи за неговото приложение и представяне на потребителите, защита от посегателство върху интелектуалната собственост и др.

Дигитализацията е процес на създаване на дигитални файлове чрез конвертиране на аналогови носители, които могат да бъдат традиционни архивни материали (документи, планове, фотографии) или триизмерни обекти като музейни артефакти или произведения на архитектурата. Като метод дигитализацията е приложение на нови подходи и средства за преобразуване на едни източници в други чрез използване на информационни и комуникационни технологии. Резултатът е дигитално копие или дигитален сурогат, който бива класифициран като дигитален носител и е обект на същите предизвикателства, свързани с опазване на достъпа до него, както и ако беше „дигитален по рождение“ (“digitally born” - материал, създаден в дигитална форма, който няма аналогов еквивалент).

Цифровите обекти са „цифрово изображение“ на различни видове обекти или съдържание (текстово съдържание, изображения, фотографии,

документи, аналогови аудио и видео записи или обемни предмети и артефакти). Цифровите обекти могат да се изградят чрез два основни типа методи:

- Създаване на ново цифрово съдържание, което няма аналогов еквивалент;
- Преобразуване на реален обект или сигнал в цифрово съдържание, където процесът на преобразуване се нарича обобщено „цифровизация” или „ДИГИТАЛИЗАЦИЯ”.

Задачите на дигитализацията могат да се синтезират в някои основни направления:

- **Запазване на фондовете и архивите** – много от дигитализираните обекти имат нетрайна или крехка структура, влияят се от атмосферните условия и от течението на времето, за тях дигитализацията е единствената надежда за съхранение;
- **Едновременно достъп до материалите** – повечето от обектите, подлежащи на дигитализация са редки и уникални предмети на историческото минало и имат безценна стойност, процесът на дигитализация ще позволи на повече заинтересовани потребители да се докоснат до тях;
- **Съхранение на фондовете в дигитални формати** – архиви, сайтове, дигитални библиотеки. Засилване на международния обмен и популяризация;
- **Предоставяне на достъп през компютърни мрежи** – лесен достъп до дигиталните архиви във време на постоянна заетост, достъп до архивите на хора с увреждания;
- **Предоставяне на нови възможности за работа с фондовете** – дигитализираните материали са достъпни до потребителите с цялата функционалност на уеб пространството да бъдат копирани, размножавани, препращани и т.н., без това да застраши целостта и здравината им;
- **Пълнотекстово търсене** – организацията в дигитални архиви способства за по-лесното откриване на търсения обект сред цялото множество, за разширено търсене, за обединение на търсени резултати;
- **Класификация на дигиталните фондове чрез метаданни** – цялото богатство на фондовете може да бъде съпровождано от важната

информация за авторски права, дата на създаване, идентификационен номер и т.н.

Дигитализацията на славянските ръкописи представлява комплексен и интердисциплинарен процес, който обединява достиженията на палеографията, кодикологията, текстологията, библиотекознанието, информатиката и съвременните цифрови технологии. Нейната основна цел не се изчерпва със създаването на висококачествени цифрови копия на ценни ръкописни паметници, а включва изграждането на цялостни цифрови представяния, които осигуряват тяхното дългосрочно съхранение, научно описание, семантично аотиране и ефективно използване в изследователската и образователната практика. По този начин се създават условия за значително разширяване на достъпа до ръкописното наследство, без да се нарушава физическата цялост на оригиналите.

Успешната реализация на процеса изисква прилагането на утвърдени международни стандарти за дигитализация, описание и обмен на данни, използването на съвременни технологии за висококачествено цифрово заснемане, обработка и архивиране на изображения, както и разработването на богати метаданни, които гарантират оперативна съвместимост, дългосрочно съхранение и повторна използваемост на цифровите ресурси. Особено значение придобива интегрирането на технологии за автоматично разпознаване на текст, семантично моделиране, свързани данни (Linked Data) и средства на изкуствения интелект, които създават нови възможности за автоматизиран анализ, търсене, интерпретация и извличане на знания от ръкописните паметници.

В този контекст дигитализацията се превръща в стратегическа предпоставка за изграждането на съвременни цифрови библиотеки и изследователски инфраструктури, които подпомагат международното научно сътрудничество, обмена на знания и развитието на цифровите хуманитарни науки. Тя не само съдейства за опазването и популяризирането на славянското писмено наследство, но и открива нови възможности за неговото интегриране в съвременната научна, образователна и културна среда. По този начин цифровите представяния на ръкописите се превръщат в активен научен ресурс, който стимулира нови интердисциплинарни изследвания и допринася за по-задълбоченото изучаване, съхраняване и популяризиране на славянската книжовна традиция в национален и международен контекст.

3. Анализ и препоръчителни методи, техники и технически средства за цифрово заснемане/сканиране с оглед осигуряване на цифрово съдържание, подходящо за анализ, както от експерти в областта, така и чрез средствата на изкуствения интелект и други видове машинен анализ

Подготовката за цифровизация е един от най-важните етапи в жизнения цикъл на цифровите културни ресурси, тъй като определя качеството, надеждността и устойчивостта на целия процес. Тя изисква внимателно планиране и координация на организационни, технологични и методически дейности, които гарантират, че създадените цифрови представяния ще отговарят на съвременните изисквания за качество, оперативна съвместимост и дългосрочно съхранение. Особено при цифровизацията на документи и ръкописи е необходимо да се отчита тяхната историческа, културна и материална специфика, както и изискванията за опазване на оригиналите по време на процеса. Ефективната подготовка включва избор на подходящи стандарти, технологии и работни процедури, осигуряване на необходимата техническа инфраструктура, както и планиране на последващото описание, управление и използване на създадените цифрови ресурси.

Подготовката за цифровизацията на обекти от културно-историческата област, в частност на документи и ръкописи, включва следните процеси:

- Планиране на работни процедури;
- Подбор на ресурсите за цифровизиране;
- Преглед на правата над интелектуалната собственост;
- Избор на хардуер и софтуер;
- Подбор на цифрови формати и стандарти;
- Планиране на цифровизацията като процес;
- Цифрова обработка на получените сурогати;
- Индексиране на цифровите ресурси - мета данни и мястото им в цялостния процес на управление на информацията;

- Дизайн на пространството за съхранение.

Въпреки значителния напредък на технологиите за цифровизация, процесът е съпътстван от редица организационни, технически и методологични предизвикателства, които могат да окажат съществено влияние върху качеството, надеждността и дългосрочната използваемост на създадените цифрови ресурси. Най-често срещаните проблеми са:

- **Качество на цифровото представяне и избор на файлов формат** – качеството на цифровите изображения зависи от множество фактори, включително използваното оборудване (скенери, фотоапарати, осветление), настройките за заснемане и последващата обработка на изображенията. Неподходящият избор на хардуер или софтуер може да доведе до загуба на детайли, изкривяване на цветовете, ниска разделителна способност или наличие на артефакти. Съществено значение има и изборът на файлов формат. За архивните копия следва да се използват устойчиви и беззагубни формати (например TIFF), докато за предоставяне на достъп чрез интернет могат да се използват компресирани формати (JPEG, JPEG2000, PNG, PDF и др.). Използването на нестандартни или слабо разпространени формати създава риск от бъдеща несъвместимост и затруднен достъп до цифровите ресурси.
- **Описание на цифровите обекти и метаданни** – цифровото изображение само по себе си не е достатъчно, за да бъде ресурсът ефективно използван. Всеки цифров обект трябва да бъде придружен от пълни, точни и стандартизирани метаданни, които описват неговия произход, съдържание, авторство, датировка, физически характеристики, права за използване и технически параметри на цифровизацията. Липсата на качествени метаданни значително затруднява откриването, идентифицирането, научното описание и повторната употреба на ресурсите, независимо от високото качество на самите цифрови изображения.
- **Предоставяне на достъп и представяне пред потребителите** – успешната цифровизация не завършва със създаването на цифровите файлове. От съществено значение е те да бъдат интегрирани в удобни за използване цифрови библиотеки, архиви или репозиториуми, които предлагат ефективни средства за търсене, разглеждане, сравнение и изтегляне на ресурсите. Лошият дизайн на потребителския интерфейс, ограничените възможности за търсене или липсата на

стандартизирани услуги могат значително да намалят практическата стойност и използваемостта на цифровата колекция.

- **Защита на интелектуалната собственост и сигурност на цифровите ресурси** – цифровите копия трябва да бъдат защитени както от неоторизиран достъп, модификация или унищожаване, така и от неправомерно използване, нарушаващо авторските и сродните права. Това налага прилагането на политики за управление на правата върху цифровото съдържание, използване на водни знаци, лицензиране, механизми за удостоверяване на автентичността, контрол на достъпа и редовно архивиране на цифровите файлове. Необходимо е също така да се гарантира дългосрочната сигурност и съхранение на цифровите ресурси чрез надеждни системи за резервиране и цифрово архивиране.
- **Дългосрочно съхранение и технологична устойчивост** – един от най-сериозните проблеми при цифровизацията е осигуряването на дългосрочна достъпност на цифровите ресурси. Бързото развитие на информационните технологии води до остаряване на файлови формати, програмни средства и носители за съхранение. Това налага периодична миграция на данните към нови формати и платформи, използване на международни стандарти и прилагане на политики за цифрово съхранение, които гарантират устойчивостта и повторната използваемост на цифровите обекти през времето.

3.1. Подходи за дигитализация

Подходите към дигитализацията се определят от вида, състоянието и предназначението на културния обект, както и от целите на цифровото представяне. При документалното и ръкописното наследство основен приоритет е постигането на максимална точност при възпроизвеждането на оригинала, като същевременно се гарантира неговото физическо опазване по време на процеса. Изборът на подходяща технология – плоско сканиране, планетарно сканиране или цифрово фотографиране – зависи от характеристиките на обекта, неговата материална основа, степента на запазване и изискванията към качеството на цифровото копие. Независимо от използвания метод, процесът следва да осигури висока разделителна способност, вярно цвето предаване, минимални геометрични изкривявания и съответствие с утвърдените международни стандарти за дигитализация на културно наследство.

Подготовката за цифровизация на изображение започва с избор и подготовка на определените документи или текстови материали на хартия (пергамент или друга основа) и с въвеждане в работен режим на системата за цифровизация.

Системата представлява – сканиращо устройство, свързано с персонален компютър със съответните софтуерни компоненти за извършване на процеса на сканиране.

Скенерите представляват набор от оптични елементи (източник на светлина, оптичен преобразувател и светочувствителни елементи), монтирани заедно в едно общо много прецизно устройство. Това тяло се придвижва по дължина на сканирания документ от механизъм или ръчно (при ръчните скенери). Източникът на светлина е лампа (най-често луминисцентна) с точно определени стойности за силата на светлината и цветната температура. Цветната температура има голямо значение за верността на цвета на копие. Качеството на системата от огледала или призма, с чиято помощ се връща отразената светлина от повърхността на документа и се насочва към светочувствителните елементи, също е много важно за крайния резултат от сканирането. Основният недостатък на този тип скенери е обстоятелството, че светлината, която попада върху светочувствителните елементи е отразена светлина – от повърхността на документа, който се сканира. Нейното количество зависи от коефициента на отражение на повърхността, която се сканира. Понякога е невъзможно да се получи използваемо копие от сканирания обект, поради много голям или много малък коефициент на отражение. Естествено, има и софтуерни филтри и математични модели за корекции на полученото копие. Тези филтри могат да се настройват преди самото сканиране или да се използват, след като е сканиран документа, като се прилагат върху файла, носещ данните от процеса на сканиране.

3.1.1. Класификация на скенерите

Изборът на подходящ скенер е един от ключовите фактори за качествената дигитализация на документално и ръкописно културно наследство. Различните видове скенери се отличават по своята конструкция, принцип на работа, разделителна способност и предназначение, което определя областите им на приложение. При цифровизацията на ценни и уникални документи изборът на сканиращо устройство трябва да отчита както необходимото качество на цифровото

изображение, така и изискванията за безопасно и неинвазивно обработване на оригиналите.

Ръчни (Hand-held) скенери

Този вид скенери приличат много на баркод четците, които се използват в магазините. Един от най-големите им недостатъци е това, че изискват много спокойна ръка – при сканиране на документ с такъв скенер дори и минимално отклонение встрани се откроява на сканираното под формата на изкривявания в образа. Поради това ръчните скенери се използват, когато има нужда от непретенциозни копия. Ако документът, който се сканира е по-широк, трябва сканирането да се извърши на няколко пъти и получените части да бъдат “залепени” една за друга с помощта на софтуер за графична редакция. Най-малките отклонения при сканирането обаче правят доста трудно и понякога невъзможно прецизното слепване. Обикновено ръчните скенери са с разделителна способност от 200 до 400 dpi и с тях се дигитализират изображения с размер до 4 инча (10 см).

Плоски настолни (Flatbed) скенери

Това са най-разпространените скенери. С тях могат да се сканират всякакви плоски обекти – книги, листове хартия и изобщо всичко, което може да се постави върху стъклото им. Тяхното устройство е сравнително просто и изключително подходящо за целта им. Сканираният обект се поставя върху стъкло. Под него, по цялата му дължина преминава устройство, наречено сканираща шейна. На нея има мощна флуоресцентна лампа, която осветява обекта – светлината се отразява от него и попада върху огледало, поставено до лампата – след това отразената от него светлината преминава през система от лещи и огледала и стига до CCD сензора⁹, който преобразува образа в електрически сигнали. За да се получи цветен образ, при някои по-стари скенери се използва т.н. трипасов¹⁰ начин

⁹ CCD линийката при скенерите е компонента, който превръща отразената от образа светлина, в електрически сигнали, които компютъра приема и обработва. CCD линийката е подобна по устройство на сензорите, които обработват изображението при фотоапаратите, но понеже тя не е матрица, а само ред от елементи се нарича линийка. В най-общ смисъл CCD линийката е редица от светлочувствителни полупроводникови фотоелементи, които превръщат изображението в цифрови сигнали.

¹⁰ Еднопасов / трипасов (с едно или три преминавания) Това определя с колко преминавания на шейната този скенер ще може да сканира един документ. Защо се налага да се правят три преминавания ? Ами CCD сензора обикновено не е цветен и за да се различават основните цветове трябва да му се прилагат филтри. При еднопасовите скенери сензора е разделен по

– при него сканиращата шейна преминава три пъти един след друг под образа, като всеки път се събира информация за един от трите основни цвята – така може да се получи по-качествено изображение, но е бавен и затова при повечето скенери днес се ползва еднопасов метод – шейната минава един път, а отразената светлина се разделя на три части чрез призма и попада върху различни части от CCD линейката – после трите получени изображения се събират софтуерно.

Плоските настолни скенери работят с разделителна способност от 300 до 1600 dpi (с интерполация), а размерът на работното поле е от A4 до A3.

На плоските настолни скенери трудно се сканират дебели книги или документи с неравна повърхност. Същото се отнася и за документи в риск за нарушаване на целостта или качеството им (като средновековни ръкописи например).

Барабанни скенери (Drum)

Барабанните скенери са насочени предимно към хората занимаващи се с предпечатна подготовка, фотообработка или чертане. Те осигуряват високодетайлен образ, с много високи резолюции (4200 dpi).

Оригиналът се поставя върху барабан. При едновременно линейно движение на оптичен сензор и въртене на барабана се снима информация за цвета и оптичната плътност на всяка точка от оригинала. С развитието на дигиталната фотография в издателската дейност все повече се използват изображения, получени директно от фотоапарата, без да се преминава през сложния процес на: снимане с широкоформатен фотоапарат – проявяване на качествена снимка – сканиране с барабанен скенер.

Широкоформатни скенери

Широкоформатните скенери се използват за сканиране на големи чертежи, в машиностроенето, архитектурното проектиране, дизайн, картография и т.н. На широкоформатни скенери може да се сканира с ширината на формати до A0 включително. Сензорът е линеен и

цялата си дължина на три участъка – по един за всеки от основните цветове. Това обаче не позволява да се постигне цялата реална резолюция на елемента. При трипасовите скенери шейната преминава три пъти, по един за всеки основен цвят.

неподвижен. Това позволява документът да се задвижва и преминава през сензора и така да няма ограничение за неговата дължина.

Слайд (Slide) скенери.

Постигат максимална разделителна способност 4200 dpi. Работното им поле е обикновено с размера на слайдовете 35x35 мм или ивици с ширина 35мм. Принципът на работа е коренно различен спрямо повечето типове скенери, обработващи обикновено отразените лъчи, тъй като те обработват пропуснатите, т.е. светлината трябва да преминава през оригинала.

Проекционни модели. Book Scanners

Проекционният скенер е безконтактен тип устройство, което служи за дигитализиране на оформени по различен начин документи. Познати са още и като планетарни или орбитални скенери поради специфичния дизайн: фотоклетките са на фиксирано разстояние от оригинала, тоест, като в орбита. Разпространено е и наименованието им „бук скенери“ (book scanners), поради широкото им използване при сканирането на книги. Този тип скенери намират все по-широко приложение в музеи, изследователски центрове и други подобни институции, където е необходимо да се работи с документи, които по правило имат специфично състояние, съхранение и използване. Предимства на този тип скенери са вградени технологии и функции за сканиране на множество документи, за стабилизиране на изображения, за автоматични корекции и ретуширане, за оптимизиране на сканираните изображения, за безконтактно обръщане на страници с времеви режими за оптимално сканиране, и т.н. По-добрите от тях разполагат и с функции за дигитални корекции на пространствени триизмерни изкривявания и деформации на оригиналите и за подобряване на заснемането на области с избледняло съдържание, като се премахват проблемите, предизвикани от естествената светлина, намаляват се до минимум ефектите на "размазването" в сканираното изображение и всички евентуални изкривявания в сканирания образ.

Използването им за сканиране на средновековни ръкописи и старопечатни книги е препоръчително.

Специализирани скенери

Специализираните скенери се различават съществено от описаните по-горе само по това, че правят копия или цифровизират информация с помощта на светлина, невидима за човешкото око. Това може да бъде инфрачервена или ултравиолетова светлина. Има скенери, които работят дори и с радиовълни, като медицинските скенери. Това се налага, защото специалните данни, които се сканират са невидими при естествена светлина. Понякога специалните скенери се конструират така, че да могат да прочетат информацията от сканирания обект, без значение в какво положение се намира той спрямо скенера. Такива, например, са скенерите за бар код, поставен на опаковките на стоките в магазина или, например, във високотехнологичен склад. Такива типове специални скенери са:

- за сваляне на отпечатьци от пръсти при системи за достъп;
- за паспортни проверки;
- за чекове;
- за триизмерно сканиране;
- медицински скенери.

Скенерите за триизмерно сканиране са по-сложни и не винаги работят по начините, описани по-горе. Има случаи, при които, за да се получи добро триизмерно копие, е достатъчно да се сканират само определени зони или точки от обекта. Те се описват в тримерна координатна система. След това, по тези точки се съставя виртуален модел на сканирания обект, а повърхността му се получава чрез компютърно генерирани материали, които имитират реалната повърхност на оригинала.

Изборът на подходящ тип скенер е определящ за качеството, ефективността и безопасността на процеса на дигитализация. Не съществува универсално решение, приложимо за всички видове културни обекти – всеки тип скенер има своите предимства, ограничения и специфични области на приложение. При дигитализацията на славянски ръкописи и други ценни документални паметници основен приоритет е съхраняването на физическата цялост на оригиналите, поради което все по-широко приложение намират безконтактните и планетарните скенери. В същото време за масова дигитализация на съвременни документи успешно се използват листоподаващи и плоски скенери.

Съвременните тенденции в развитието на сканиращите технологии са насочени към повишаване на качеството на цифровите изображения,

автоматизиране на процесите, използване на интелигентни алгоритми за обработка на изображенията и интегриране на технологии, базирани на изкуствен интелект. Това създава предпоставки за по-прецизна цифровизация, по-богато семантично описание и по-ефективно използване на цифровите ресурси в научните изследвания, цифровите библиотеки и системите за дългосрочно съхранение на културното наследство.

3.1.2. Показатели за оценка на скенери

Оценката на характеристиките на скенерите е съществен елемент при избора на подходящо устройство за дигитализация на документи и културно-исторически обекти. Тя се основава на набор от технически и експлоатационни показатели, които определят качеството на цифровото изображение, производителността на процеса и пригодността на скенера за конкретни приложения. Сред най-важните показатели са оптичната разделителна способност, цветната дълбочина, динамичният диапазон, точността на цветопредаването, скоростта на сканиране, максималният размер на обработваните документи, възможностите за безконтактно сканиране и автоматизирана обработка на изображенията и др. Правилната оценка на тези характеристики позволява избор на оптимално техническо решение, съобразено с изискванията за качество, дългосрочно съхранение и последващо използване на цифровите ресурси.

Резолюция. Резолюцията е един от най-важните показатели за скенерите. Тя показва доколко е качествено изображението, което се получава след сканиране. Измерва се в dpi (dots per inch – точки на инч) . Резолюцията при скенерите е показател за това, по колко отделни точки може да разпознае един скенер на определено разстояние върху образа, който се сканира. Обикновено за скенерите резолюцията се показва с две цифри – примерно 300x300 dpi – така се показва за един квадратен инч, колко точки се разпознават по хоризонталата и вертикалата.

Разделителната способност се дели на: оптична и апаратна. Колкото е по-голяма оптичната разделителна способност, толкова повече светлочувствителни елементи притежава скенерът. В момента за нормална разделителна способност, дори в по-ниския клас скенери, се счита 600 dpi.

Апаратната разделителна способност (по вертикала) се определя от възможния брой стъпки на механиката за придвижване на оптичните елементи за един инч.

Видът на оптичната система зависи единствено от вида на фоточувствителните елементи. Тя може да бъде CCD (Charged Coupled Devices) или CIS (Compact Image Sensor). Скенерите, използващи CCD, дават по-качествени изображения, но оптичната им система е доста сложна, с големи размери и сериозен разход на енергия. Скенерите, използващи CIS, имат доста опростена оптична система, компактна и лека, а разходът на енергия не е много голям. Основният недостатък е липсата на много голям контраст. Примерни резолуции, достатъчни за някои типове документи:

- Черно бяло копие на текстов документ – от 100 до 150 dpi;
- Цветно копие на текстов документ – от 150 до 200 dpi;
- Копие за Интернет – от 72 до 150 dpi;
- Копие за печат или вграждане в текстов документ от 150 до 300 dpi;
- Копие за брошура или друг печатен рекламен материал - минимум 600 dpi.
- Копие на ръкописни и старопечатни документи - минимум 600 dpi, като е препоръчително да е с резолуция 1200 dpi.

Ако е необходимо копие на снимка или документ за широкоформатен печат, добри резултати се получават само, ако копие е с резолуция поне 300 dpi, като е препоръчително да е 600 dpi, и мащаб минимум 1:10.

Дълбочина на цвета. Важен момент, който трябва да се следи, е дълбочината на цвета. Известен още като дълбочина на битовете, този показател се отнася към количеството информация, която може да бъде обхваната от всеки отделен пиксел при сканирането. На практика, колкото е по-висок този показател, толкова по-добре устройството ще определя леките разлики в цветовете. Стандартно, скенерите използват 24 бита цвят (по 8 бита за всеки един основен цвят). Някои скенери обаче имат възможност за поддържане на 48 бита цвят, което дава възможност за коригиране на цветовете (при последващи обработки на копие), без загуба на информация. Това е така, защото с колкото повече бита е описан един цвят (а за компютрите всеки нюанс се разглежда като нов цвят) толкова по-верен е той спрямо оригинала.

Динамичен обхват, т.н. D-max (съкращение на Maximum Image Density), т.е. максимална плътност на изображението (колкото по-плътна е едно изображение, толкова по-тъмно става).

Технология. Въпреки, че не е от критична важност, трябва да се знае какъв тип сензор използва скенерът. Най-старата технология, която и в момента се употребява, е CCD (Charge Coupled Device) - светлинно-сензорна система, която се използва и при цифровите камери. Тази система е низ от светлочувствителни диоди, наречени PHOTOSITES, които генерират електрически сигнал при допир със светлинни фотони. Всеки от тези PHOTOSITES записва пиксел, малка точка представлява светлината, която попада в тази точка.

По-нова технология е CIS (Contact Image Sensor – контактен сензор), която макар и да не осигурява качеството на CCD, навлезе в употреба благодарение на много ниската си консумация на енергия. Това означава, че скенерът би могъл да работи и през USB порт, без да е необходимо да се захранва от електрическата мрежа.

3.1.3. Интерфейс

Интерфейсите за скенери са фундаментален хардуерен и софтуерен компонент, който осигурява двупосочна комуникация между сканиращото устройство и компютърната система. Те определят както физическата връзка, така и скоростта на предаване на данните и съвместимостта със стандартите за дигитализиране на изображения и документи.

Връзката (интерфейс) на скенера с компютърната система определя единствено и само скоростта, с която става предаването на информация от скенера към компютъра. Връзката е двупосочна. С помощта на специален диалог помежду си двете устройства следят целия процес на предаване и приемане на данни. Когато настъпи забавяне или грешка при предаването, процесът на сканиране може да бъде прекъснат и продължен по-късно, като при нужда част от оригинала може да се сканира отново, като накрая се получава едно цялостно копие.

Срещат се три варианта на свързване към компютъра – чрез паралелен порт, чрез USB и SCSI интерфейса. Някой от старите скенери работят с паралелен интерфейс, но тъй като все повече системи поддържат USB портове, USB скенерите вече изместиха паралелните. На теория скоростта на USB порта е между тази на по-бавния паралелен порт и бързата SCSI връзка. Но докато скенер с SCSI интерфейс изисква добавяне на SCSI адаптер, да се зададат адреси и да се терминира последното устройство, то инсталирането на USB скенер става съвсем лесно и бързо.

Друг фактор, освен интерфейса, определящ скоростта на въвеждане на информацията, е разделителната способност, с която се работи по време на сканирането. Трябва да се знае за какво ще се използва сканираното изображение и според това да се определи разделителната способност, с която ще се сканира. **Ако за някакви цели са необходими копия на един оригинал с различна разделителна способност, е добре оригиналът да се сканира веднъж с най-голямата необходима и възможна разделителна способност.** След това на базата на конвертиране на формати и оптимизации от първоначалното дигитално копие може да се създадат необходимите версии за ползване за конкретни цели.

По отношение на средновековни ръкописи и старопечатни книги е препоръчително да се сканира поне с резолюция 600 dpi.

4. Технически характеристики за високо качество на цифровото изображение (резолюция, дълбочина на пикселите, цветови модел и формат за съхранение на файлове и др.)

Резултатът от цифровизирането на артефакти, в частност средновековни ръкописи и старопечатни книги, са специфични компютърни обекти, чиято структура и метаданни отговарят на строги технически характеристики. Тези параметри варират фундаментално в зависимост от физическия носител на оригиналния източник и целите на дигиталния архив. В научната практика класификацията на файловете формати се извършва въз основа на доминиращия тип информация, която те капсулират: статични двуизмерни изображения (растерна графика), структуриран текст, векторни компоненти или хибридни масиви.

При дългосрочното планиране и реализация на процесите по дигитализация критично значение имат следните пет технологични и методологични стълба:

Качество и автентичност на дигиталното съдържание: То се определя от хроматичната точност (цветово съответствие), пространствената разделителна способност (DPI/PPI) и липсата на артефакти от компресия. За архивни цели (master copies) се изисква съхранение в некомпресирани формати без загуба на данни (напр. TIFF, RAW), които капсулират максимален обем информация от сензора.

Методологически подходи за дигитализация: Изборът между безконтактно планетарно сканиране, дигитална макрофотография или специализирани V-образни люлки за деликатни кодекси е водещ. Методът определя както физическата безопасност на артефакта, така и геометричната коректност на репродуцираното дигитално изображение.

Интерполация и съвместимост на файловете формати: Наложително е използването на отворени, стандартизирани и устойчиви във времето формати (като TIFF за изображения, PDF/A за документи и XML/TEI за текстова транскрипция). Това предотвратява риска от софтуерно остаряване (software obsolescence) и гарантира четимост от бъдещи информационни системи.

Капацитет за мрежово споделяне и оперативна съвместимост: Интеграцията на протоколи като IIF (International Image Interoperability

Framework) позволява бързо и детайлно уеб-базирано визуализиране на тежки графични файлове. Оптимизацията на производни формати (напр. JPEG 2000 или WebP) осигурява баланс между детайлност на изображението и нисък мрежов трафик.

Мултиплатформеност и архитектурна независимост: Дигиталните ресурси трябва да бъдат напълно достъпни и функционални независимо от операционната система (Windows, macOS, Linux/Unix) или хардуерната архитектура на крайния потребител. Това изисква софтуерно-независимо кодиране на метаданните и придържане към международните уеб стандарти на W3C.

Пълнотата на статично изображение се измерва в брой пиксели, резолюция, брой цветове и компресиране на изображението. Следните параметри за запазване на цифрово изображение с високо качество могат да се използват: разделителна способност – от 600 dpi до 2000 dpi; брой цветове – над 16 битов цвят за всеки канал от RGB схемата; файлови формати: TIFF или JPG (с минимална компресия).

4.1. Описание на форматите за съхранение на изображения

Дигитализираните графични изображения могат да се разделят на няколко групи в съответствие с принципите, по които са изградени. В тази връзка се оформят два основни типа изображение – растрни и векторни. И двата типа могат да се записват във файл в натурален или в компресиран вид. Компресирането се налага от съображения за икономия на памет и дисково пространство.

4.1.1. Векторни графични изображения и формати

Векторните графични изображения са тези, при които описанието на графичната информация се представя във вид на векторни множества или зависимости. Те са удобни за представяне на изображения, които могат да се редуцират до математически криви. Предимството по отношение на памет е безспорно. За сравнение, ако е необходимо да се изобрази квадрат с начални координати 0,0 и размери 100 пиксела, за неговото представяне в

растерен вид са необходими 10 000 байта, при положение, че 1 пиксел се представя в 1 байт. Същото изображение може да се представи във векторен вид с помощта на 9-16 байта – по два (четири) байта за координатите всеки връх на квадрата и един (три) байта за представяне на цвета.

4.1.2. Растерни графични изображения и формати

Растерните изображения са тези, при които представянето на графичната информация се осъществява по алгоритъм, при който на всеки един пиксел от изображението съответстват от 1 бит до 4 байта. Основен параметър при тези изображения е разделителната способност. Тя се представя като съотношение на графични единици към линейни единици, като най-често използваната мерна единица е „dpi”. Нормално е да се работи и в пиксели за милиметър. До голяма степен размерът на едно графично изображение зависи от разделителната способност, с която се работи.

5. Препоръчителен софтуер и настройки. Работата с избран софтуер за цифрова обработка

Проблемите, свързани с дигитализацията основно се отнасят до качеството на файла и формата, използван за съхранението му – понякога изборът на неподходящ за нуждите хардуер или софтуер води до намаляване в качеството на дигитализираните обекти. Форматът също е от особено значение, той трябва да бъде поддържан и разчетен от най-разпространените програмни средства, в противен случай се рискува обектите изобщо да не достигнат до крайните потребители, за които всъщност са предназначени. Описанието на файла на документа е важно може би колкото самия документ, тъй като без съпровождаща или неточна информация обектът е трудно неизползваем. Адекватното и ясно представяне на дигиталния архив пред потребителите също играе много важна роля.

Основните видове обработки, на които се налага да бъдат подлагани файловете в процеса на дигитализация са следните.:

- Отстраняване на изкривявания и на дефекти;
- Разширяване диапазона на яркостта;
- Отделяне контурите на изображението;
- Оцветяване черно-белите части на изображението;
- Корекция на недостатъци, получени при експониране на отделни части;
- Въвеждане на нови детайли;
- Промяна на формата и размера на съставляващите компоненти;
- Разпознаване на изображения;
- Сравняване на изображения (Pattern recognition) – Техника за сравняване на съществуващи изображения с изображения от предварително определени категории, която използва различни статистически методи.

От решаващо значение за решаване на проблемите, свързани с дигитализацията, и за извършване на съответните обработки е изборът на софтуер.

5.1. Софтуер

Почти всички скенери се доставят с основен софтуер, който е необходим при сканиране. Използването му, особено за постигане на високо качество на дигиталните копия, не винаги е препоръчително. Скенерите от по-висок клас използват и специализиран софтуер за сканиране и обработка на изображения.

Всички програми за цветно сканиране предлагат някои функции с помощта на които може да се определят следните етапи на сканиране:

- Избор на изображение и подготовката му за сканиране – почистване от драскотини, прах, следи от пръсти.
- Оценка на оригинала – базира се на следните фактори :
 - експозиция – дали оригиналът е преекспониран или недоекспониран (той е по-тъмен и не улавя всички подробности);
 - цветен оттенък (преобладаващ цвят);
 - рязкост – тази част от изображението, което е на фокус.
- Предварително сканиране (prescan) – това е бързо сканиране при ниска разделителна способност, което позволява да получите предварителен изглед на изображението и да нагласите размерите му (cropping). Ако скенерът и софтуерът позволяват може да се направят някои настройки и корекция на цветовете. Следва корекция в рязкостта (sharpening) – когато е необходимо да се коригира замъглен оригинал или ако при сканирането оригиналът не е бил фокусиран добре.
 - Изрязване;
 - Управление на разделителната способност;
 - Възможност за избор на типа на оригинала - цветен, степен на сивото или шрих;
 - Управление на яркостта и контраста;
 - Баланс на цветовете;
 - Подобряване на рязкостта;
 - Установяване на размера на изображението и разделителната способност;
 - Преобразуване от негатив в позитив (когато е необходимо);
 - Записване на изображението в подходящ файлов формат.

5.2. Софтуерни изисквания – програмни средства за дигитализиране на изображения

Изборът на софтуер трябва да е съобразен с типа на материалите, които се дигитализират. От софтуера се изисква да обработва дигиталния резултат в посока корекция на цветовете, подрязване на краища или компресиране на файла с цел представянето му в уеб. Има много такива платени или безплатни програми и изборът зависи конкретно от целите и възможностите на проектите.

Софтуерът трябва да може:

- да интерпретира много големи файлове;
- да модифицира резолюцията и дълбочината на цвета;
- да съхранява различни версии с различни големина на файла;
- да копира част от образа и да го съхранява като отделен файл;
- да експортира образи в различни файлови формати.

След процесът на сканиране, полученото изображение трябва да се съхрани във формати, подходящ за организиране в архив и/или за показване в уеб, което предполага допълнителни обработки.

5.3. Дигитализация и обработка на изображения. Формати за съхранение на дигитализираните обекти

Компютърната графика е закономерен резултат от развитието на компютърните технологии. В съвременните компютърни системи всички изразни средства, от диалога с потребителя, през контрола на изчислителния процес до визуализацията на резултатите се постигат със средствата на компютърната графика.

Обработка на изображение (Image Processing) е процес на създаване, анализ, преобразуване или интерпретация на изображения:

1. Въвеждане на изображението с помощта на скенер или цифров фотоапарат, под управлението на графична програма.
2. Компресиране на изображението, когато се изисква голям капацитет памет за съхранение или е нужно улесняване на достъпа.
3. Обработка и спецификация на формата на файла, който ще е най-подходящ за съхранение.

Изображения биват преобразувани в компютърни с помощта на входните графични устройства - скенери, и др., и с помощта на компютър се осъществява преобразуването на изображенията. За тази цел се използват програмни продукти като Adobe Photoshop, Photo Paint и др. Те предлагат обширно многообразие от пособия за първоначална обработка на изображения, като коригиране на яркостта, контраста, цветовата гама, разпознаване на контурите на обектите и др. Съществуват и безплатни версии (с намалени функционалности) на платените програмни продукти, както и безплатни програми (например GIMP) за обработки на изображенията. През последните години се появиха и онлайн версии на програмни продукти или изцяло онлайн редактори на изображения (например Photopea), които при отсъствие на проектни възможности може да бъдат използвани в процеса на обработката на изображения.

5.4. Описание на форматите за съхранение и представяне на изображения в уеб (JPEG, GIF, PNG, DJVU)

Ефективната визуализация на графична информация в глобалната мрежа изисква прецизен баланс между естетическото качество на изображението и техническата оптимизация на уеб ресурсите. За разлика от архивните системи за дългосрочно съхранение (където водещ критерий е абсолютната автентичност и липсата на загуби), при представянето на изображения в уеб среда ключов фактор е скоростта на зареждане на страниците. Обемът на графичните данни оказва пряко влияние върху мрежовия трафик, производителността на клиентските устройства и цялостното потребителско изживяване (User Experience - UX).

Изборът на конкретен уеб формат се определя от характера на визуалното съдържание (фотореалистични сцени, графични интерфейси, текст, лога или анимации) и поддържаните от него функционалности. В тази връзка съвременната уеб разработка класифицира форматите според алгоритмите им за компресия, поддръжката на алфа-канал (прозрачност) и възможността за адаптивно скалиране.

Настоящата подчасть предлага систематизиран анализ на утвърдените стандарти и съвременните технологични алтернативи за представяне на изображения в Интернет.

5.4.1. Файлов формат JPEG

JPEG е често използван метод за компресия на растерни изображения. Включва компресия със загуба на данни, така че позволява намаляване на размера на файловете за сметка на намаляване на качеството на изображението. Поради спецификата на използвания метод за компресия, JPEG е по-подходящ за многоцветни изображения, отколкото за такива, съдържащи големи едноцветни области.

При JPEG компресията се използва т.нар. метод за компресиране със загуби. Чрез такъв вид метод изображенията стават с доста по-компактен обем. Загубите се изразяват в това, че част от началната информация на изображението се загубва при компресирания файл, което води до понижаване на качеството на изображението при възстановяването му. Твърде високите компресии водят до такава загуба на информация, че възстановеното изображение е неизползваемо.

Името е съкращение от Joint Photographic Experts Group, името на комитета, който създава стандарта през 1992 г. Наричан JPEG или JPG стандартът поддържа до 16,8 млн. цвята (24-битов цвят). През 1994 г. е одобрен като международен стандарт ISO 10918-1. JPEG стандартът представлява кодек, което означава, че изображението се компресира в поток от байтове, а след това обратно се декомпресира в изображение. Обикновено при компресирането чрез „компресия със загуби“ съзнателно се губи определена част от качеството на изображението (намаляване на броя цветове и др.), за да се постигне значително намаляване на неговия размер. Но също така има и различни вариации, при които се намалява до минимум загубата на качество. При компресиране може да се настройва степента на загубите.

Използваните разширения на такива компресирани файлове са .jpg и .jpeg, както и .jpe, .jfif и .jif. От тях .jpg е най-разпространеното.

5.4.2. Файлов формат GIF

GIF (Graphics Interchange Format) е разработен от CompuServe. Първоначално той е бил създаден за улесняване на трансфера на графичните изображения по мрежата за онлайн услуги на тази компания. След това се популяризира и сега има важна роля за уеб. Използва алгоритъм за компресия LZW (Lempel-Ziv&Welch). Това е тип компресия

без загуби, т.е. запазва всички първоначални данни по време на процесите на компресия и декомпресия. GIF поддържа прозрачност (transparency) на определен цвят в графичното изображение. Форматът е използван широко в Интернет, когато трябва да се съхрани изображението на лого или малка анимация. Цветовата палитра на пикселите в едно изображение в GIF е ограничен до 256 различни цвята или по-малко (до 64 градации на сивото) т.е. за описание на цвета се използват 8 бита. Той е подходящ за опростени изображения с малко цветови вариации - което прави този формат неподходящ за съхраняване на многоцветни изображения. Разширението на името е .gif.

5.4.3. Файлов формат PNG

Portable Network Graphic (PNG) е създаден с идеята да подобри формата GIF. В PNG форматът успешно се съчетават най-добрите страни на GIF и JPG. PNG се поддава малко по-добре на компресиране от формата GIF като същевременно поддържа милиони цветове подобно на JPG формата. За разлика от JPG обаче, компресията на PNG е без загуби. Подобно на GIF, PNG може да се зарежда „прогресивно“, оставяйки у потребителя впечатление за по-бързо появяване на изображението. PNG също поддържа многостепенна прозрачност на цветовете (до 256 нива) и възможност за променяне на яркостта на изображението. Тези особености позволяват изглаждане на назъбени контурни линии. Едно предимство на PNG файловете е възможността за обявяване на прозрачни и полупрозрачни зони. Въпреки, че PNG файловете се разпознават и визуализират от почти всички браузъри, то свойството прозрачност се игнорира за някои от тях, което трябва да се вземе под внимание.

5.4.4. Файлов формат DJVU

DjVu е компютърен файлов формат, разработен през 1996 г. от AT&T Labs най-вече за сканирани изображения с текст и рисунки. DjVu е отворен стандарт. Спецификациите му, както и отворения код на приложенията на декодера (и част от самия декодер) са свободни. Форматът използва:

- разделяне на изображението на слоеве (преден с текст и фон (слой));
- последователно зареждане на страницата;

- аритметично кодиране;
- компресия със загуби за двуцветни изображения.

Това позволява да се получат висококачествени, четими изображения при минимален размер на файла, за да се публикуват лесно в Интернет.

DjVu е представяно като алтернатива на PDF и се представя добре при кодирането на повечето сканирани изображения. Цветни страници от книга се кодират във файлове по 40-70 KB, черно-бели технически документи се компресират до 15-40 KB, и древни ръкописи - до около 100 KB. Всички тези примери са доста по-добри като размер от типичните 500 KB нужни за задоволителен JPG файл. Също като PDF, DjVu може да съдържа разпознат текст, което позволява лесно копиране, поставяне и търсене на текст. За цветни изображения, които съдържат текст и картини, DjVu файловете са от 5 до 10 пъти по-малки от JPG с почти същото качество. За черно бели страници, DjVu файловете са от 10 до 20 пъти по-малки от JPG и пет пъти по-малки от GIF. DjVu файловете са също около 3 до 8 пъти по-малки от черно-белите PDF файлове от сканирани документи (сканираните в цвят документи са неподходящи за PDF).

DjVu позволява дистрибуцията в Интернет на изображения с висока резолюция, сканирани документи, дигитални документи, и фотографии. Форматът дава възможност за сканиране на цветни изображения с висока резолюция от книги, списания, каталози, ръководства, вестници, исторически или архивни документи, и правенето достъпни в уеб пространството.

Една от главните технологии зад DjVu е способността да отделя изображението на фонов слой (текстури или картини) и преден слой (текст и линейни изображения). Чрез отделянето на текстовете от фона, DjVu може да запази текста във висока резолюция (следователно така се запазват острите ръбчета и се увеличава четимостта), докато в същото време компресира фоните и картините на по-ниска резолюция.

5.5. Описание на форматите за съхранение на изображения в архив (PDF, BMP, TIFF)

Документите, предназначени за архив могат да се съхраняват и в някои от гореспоменатите формати като JPG и PNG.

5.5.1. Файлов формат PDF

Portable Document Format (PDF) е файлов формат, създаден от Adobe Systems през 1993 г. за обмяна на документи. PDF е използван за представяне на двуизмерни документи, достъпни независимо от софтуер, хардуер и операционна система. Формат с фиксирано оформление, което запазва форматирането на документа и разрешава споделянето на файла. Форматът гарантира, че когато файлът се преглежда онлайн или се отпечатва, той запазва точно предвидения от автора формат и данните не могат да се променят.

Всеки PDF файл представя пълно описание на даден двуизмерен документ (при Acrobat 3-D се поддържат и тримерни файлове), което включва текст, шрифтове, изображения, и двуизмерни векторни графики.

PDF официално е обявен за отворен стандарт на 1 юли 2008 г., и е публикуван от Международната организация по стандартизация като ISO 32000-1:2008.

5.5.2. Файлов формат BMP

Файловият формат BMP (от англ. Bitmap – карта на битовете, побитова карта) е стандартен формат на растерни изображения в Windows. Изображението в BMP формат представлява група от пиксели. Всички пиксели са с еднакви размери, но могат да имат различен цвят. Има възможност за промяна на размера на изображението в BMP формат, като се разтегля или свива, но тези действия обикновено водят до поява на нежелани ефекти като назъбване на контурите или неочаквани петна по изображението.

Обикновено файловете в BMP формати се записват в некомпесиран вид, но те могат да се компресират добре с помощта на стандартни архивиращи програми (като WinZip). BMP поддържа RGB модела. BMP е подразбиращия се формат на MicrosoftPaint, която е включена в операционната система Windows като спомагателно средство за рисуване, промяна или показване на изображения.

5.5.3. Файлов формат TIFF

Tagged Image File Format (TIFF) представлява гъвкав растерен графичен формат, който се поддържа от почти всички приложения за импортиране, експортиране или създаване на графични изображения. Това е графичен стандарт, който е независим от компютърната платформа и използваната операционна система.

Този формат също използва алгоритъм за компресия LZW (Lampel-Ziv&Welch), който управлява много добре голямо разнообразие от повтарящи се образци (Patterns) и е приемливо бърз за компресия и декомпресия на графични файлове. Последната версия е TIFF 6.0 (1992) при следните технически характеристики: до 24 бита цвят и разделителна способност без ограничения. Поддържа се и опция за компресия на графичната информация.

Форматът се поддържа широко от приложения за обработка на изображения (Adobe Photoshop и много други), от приложения за подготовка на публикации и за оформление на страници (QuarkXPress, Adobe InDesign и други), както и от сканиране, изпращане на факс, текстообработка, оптично разпознаване на символи и други приложения.

6. Особености на използваните файлови формати за цифрови колекции от славянски ръкописи и техники за цифровата обработка на изображенията

Сравнението на конкретните детайли, предимствата и недостатъци на различните файлове с изображения може да помогне да избора нужния Ви формат. Това касае:

- **Размер.** Преоразмеряването може да компрометира разделителната способност на някои файлове с изображения, а на други не.
- **Разделителна способност.** Много файлове с изображения имат минимална и максимална плътност на пикселите, като и двете трябва да са подходящи за конкретните случаи на употреба.
- **Формат.** Някои изображения изискват специален софтуер за отваряне и редактиране на файлове.
- **Качество.** Форматите на изображения със загуби или без загуби могат да повлияят на качеството на изображението при компресиране.

Основните препоръчителни формати при дигитализация на цифрови колекции от славянски ръкописи са TIFF, JPEG и PDF.

6.1. TIFF спрямо JPEG файлове

Има няколко важни разлики между JPEG и TIFF файловете. Къде или дали ще ги използвате зависи от Вашите работни потоци и изисквания. Основната разлика е компресирането при JPEG файловете, което означава, че те обикновено са много по-малки по размер от TIFF файловете. Това прави JPEG файловете предпочитана опция, когато мястото за съхранение стои на първо място. TIFF файловете обаче съхраняват много повече данни за изображенията от техните JPEG аналози. По-високата разделителна способност обаче изисква много по-големи размери на файловете, което означава, че те могат да бъдат неудобни за използване и по-трудни за съхранение.

Форматът JPEG използва компресия със загуби, което означава, че тези файлове намаляват качеството на картината, за да постигнат по-малък и по-управляем размер на файла. Това прави JPEG файловете лесни за съхраняване на устройство или преместване между облачни услуги.

Въпреки това, по време на компресиране част от оригиналните данни за изображението се губят, което води до забележими загуби в разделителната способност на изображението.

Подобно на JPEG TIFF са растерни графични файлове, но TIFF форматът използва компресия без загуби, за да запази данните за изображението. Това означава, че по-принцип TIFF файловете са големи и обикновено заемат много дисково пространство. Въпреки това, високото им качество ги прави основен избор за извършване на всякакви работи по цифрово редактиране, особено ако ги съхранявате в резервния си диск след това.

TIFF е много подходящ като изходно изображение – оригиналният файл може да се запази като архивно копие. По този начин изображението може да се съхранява с възможно най-добро качество за редактиране (в работни версии) на по-късен етап. Това е особено важно по отношение на документи (ръкописи, редки книги, и др.), до чиито оригинали поради различни причини повторното дигитализиране може да е затруднено или невъзможно.

JPEG върши по-добра работа като финализирано изображение, готово за експортиране, например когато трябва да се публикува изображението на документ на уеб сайт (галерия, библиотека, уеб архив). В сравнение с TIFF JPEG файловете не съхраняват толкова много данни за изображението, което означава, че не предлагат толкова голяма гъвкавост в процеса на редактиране.

Един от недостатъците на компресирането на изображението в JPEG е появата на „артефакти“, които представляват пикселизация или блоков вид на елементите на изображението, когато даден файл загуби твърде много данни. Има доста примери за лошо компресирани изображения на уеб сайтове или много тъмни снимки (включително в онлайн архиви и библиотеки). TIFF файловете, като файл без загуби, не произвеждат артефакти.

JPEG файловете обаче, за разлика от TIFF, са предпочитани при дизайн на уеб сайтове, тъй като малкият им размер ги прави лесни за качване и управление. Те са достатъчно малки, за да не повлияят негативно на времето за зареждане на страницата, което може да повлияе на цялостното представяне на уеб сайта.

JPEG файловете са универсално съвместими с повечето стандартни операционни системи, програми за редактиране и принтери. TIFF

файловете също са, но големият им размер може да ги направи несъвместими с някои принтери и скенери. Сканирането в TIFF файлове може да изисква по-високотехнологичен хардуер (скенери, компютри) и специализирани програми. Големината на TIFF файловете може да бъде фактор само при отсъствието на достатъчно дисково пространство, тъй като една страница може да бъде сканирана в TIFF файл от 100 MB или JPG файл от 10 MB, както и изображение от 100 MB на TIFF файл при конвертиране и компресиране в JPEG може да достигне до 500 KB (половин мегабайт). При 100 сканирани страници в TIFF обемът на документа може да достигне 10 GB. Същевременно работата с TIFF файлове изисква и по-високотехнологично ниво на десктоп конфигурации за обратотката на големите TIFF файлове.

Програмите за преглед на изображения по подразбиране на Windows и macOS могат да отворят и двата файлови формата. Цифровите редактори на изображения като Adobe Photoshop имат възможност за редактиране и манипулиране на оригиналното изображение и в двата формата.

Въз основа на сравнението между възможностите, които предоставят TIFF и JPEG, препоръчително е сканирането на средновековни ръкописи и старопечатни книги да се извършва в TIFF формат с максимално висока резолюция (поне 600 dpi) за създаването на архивна база. Тези файлове могат впоследствие с оглед на използването им да бъдат конвертирани в други формати за целите на работа с тях или за онлайн достъп до документите от архива (например в JPEG, PDF или други формати). При техническа невъзможност това да бъде реализирано, сканиране в JPG с резолюция от поне 600 dpi би могло да има задоволителни резултати.

6.2. JPEG спрямо PDF файлове

JPEG е един от най-широко използваните в света файлови формати. Способни са да компресират впечатляващи детайли в размер на файл за споделяне и са най-полезни при онлайн съхраняването и достъпа на цифрови изображения.

Основното предимство е на PDF форматът е, че това е универсален формат – което означава, че съдържанието на PDF файловете винаги ще се показва еднакво на всяко устройство. PDF форматът запазва всички компоненти на страницата непокътнати и запазва качеството на изображението при уголемяване.

Както бе отбелязано, основната употреба на JPEG е за изображения, докато PDF файловете могат да обработват всякакви текстови и базирани на изображения документи. Тъй като JPEG файловете компресират данни за изображения, размерите на файловете им обикновено са сравнително малки. Това означава, че те представляват удобен начин за споделяне на цифрови изображения в мрежата. Но JPEG файловете използват компресия със загуби, което означава, че всеки път, когато файлът се записва, се губят малко данни за изображението, което може да повлияе на качеството.

Следователно многоцелевият PDF може да бъде предпочитаният тип файл за изображения, защото предлага функция за компресиране без да нарушава оригиналното качество на файла, който се споделя. Въпреки това PDF файловете са трудни за интегриране в уеб страница, ако това е основната им цел и в този случай JPEG файловете са предпочитания избор.

Основното предимство на PDF формата е, че запазва оформлението на съдържанието, като оставя различни части на документа отворени за редактиране (с Adobe Acrobat Pro и други подобни програми като Nitro PDF Professional, Foxit Advanced PDF и други платени или безплатни десктоп или онлайн приложения). Обратно на това, JPEG файловете не могат да бъдат разделени на компоненти, защото компресират цялото съдържание в един слой. Това означава, че от двата формата само PDF позволява на потребителите да копират избран текст, който трябва да използват другаде. По отношение на сканирането и запазването на средновековни ръкописи и старопечатни книги това предимство на PDF формата не може да се използва, тъй като AAR (както и всички останали PDF редактори) не разполага с инструменти за оптичното разпознаване на символи (OCR – optical character recognition) на такъв тип текстове.

Същевременно конвертирането между PDF и JPEG и в двете посоки е бързо и лесно. Същото се отнася и до конвертирането между TIFF и PDF файловете и при нужда съхраняваните в архива TIFF файлове на сканираните средновековни ръкописи и старопечатни книги могат да бъдат използвани и в PDF формат.

7. Архивиране и представяне на цифрови колекции от славянски ръкописи

Сканирането на документи, като първа стъпка от дигитализацията им, предполага създаването на архив на получените в процеса файлове. Изискванията към такъв архив обикновено са основани на типовете документи в него, като: периодични издания, книги, карти, планове, чертежи и др., в случая – славянски ръкописи.

Като параметри за дигитализиране е добре да се използват:

- резолюция – използва се минимално 600 dpi, или по-висока резолюция по преценка според конкретния случай;
- цветна дълбочина – с оглед запазване оригиналния изглед на документа и улесняване визуалното му възприемане от потребителя печатни материали до 1944-45 г. се дигитализират в цвят, 24-bit RGB. Възможно е, и желателно за средновековни ръкописи, когато това е възможно, и сканиране в 48-bit RGB в TIFF файлове. При по-нови (съвременни) документи се преценява – ако съдържат цветни изображения, графики, диаграми и т.н., то се дигитализират с параметрите, описани по-горе (24-bit RGB). Ако не е необходимо запазване на цвета, то се използва 8-bit Greyscale.
- файлов формат – дигиталните обекти се запазват в архивен формат – TIFF без компресия.

7.1. Обработка на дигитализираните документи

Независимо от вида на документа всяка отделна страница се запазва в самостоятелен файл, вкл. кориците, ако има такива. Дигитализира се и гърба на страница, дори и когато него има някакъв вид информация като текст, номерация, изображение, печат, ръкописни знаци и т.н. Изображението се изрязва по ръба на страницата, като се следи да няма загуба на съществена информация, особено при подвързаните оригинали (книги, периодика). Ако подвързията е прекалено стегната и част от текста се губи в сгъвката, изображението се изрязва по такъв начин, че в него да остане малка част (1-2 см) от съседната страница. Когато страниците не са правилно изрязани (под прав ъгъл) изображението на страницата трябва да отразява изцяло оригинала без изрязване на нейни части.

Всички страници трябва да бъдат вертикално подравнени (изправени).

Всички файлове се конвертират от TIFF в JPEG (или PDF) за използване в онлайн архиви и библиотеки с прилагане на компресия и запазване на качество на изображението според целите на публичния достъп. Ако файловете в архива са в JPEG с висока резолюция, се извършва компресиране в нови JPEG (или PDF) файлове, като оригиналните архивни файлове се запазват.

7.2. Комплектуване на файловете

От значение за правилната работа на дигиталния архив е коректното комплектуване на файловете на дигиталните обекти:

- за ръкописи, книги, сборници и пр.: всяка книга се съхранява в самостоятелна папка (когато има случай на подвързани няколко книги в общо тяло, но информационно различни, несвързани помежду си, то всяка от тях трябва да бъде съхранявана в отделна папка);
- за листови материали: съхраняват се подходящо именувани, в обща папка за типа документи (ако между тях има такива, които попадат в серия, поредица или са свързани по някакъв друг признак, то те се обособяват в самостоятелна подпапка).
- за периодика: всяка година/годишнина от дадено заглавие се съхранява в самостоятелна папка, в нея страниците/файловете на всеки отделен брой се съхраняват в отделна папка, подходящо именувани .

7.3. Именуване и подреждане на файловете

Всички файлове трябва да получат имена по определена конвенция (схема).

Примерен вариант за ръкописи:

- Име на папка: [град]-[институция]-М-[година с 4 цифри]-[код – инвентарен номер или др.]-[име на ръкописа или разпознаваема част от него].

- Име на файл: [град]-[институция]-М-[година с 4 цифри]-[код – инвентарен номер или др.]-[име на ръкописа или разпознаваема част от него]_[брояч].tiff

- [брояч]: 001, 002, 003..... (поредност на страници).

- М – manuscripts, сигнатура за визуално разпознаване на папки и файлове на дигитализирани ръкописи.

- В папката със страниците на ръкописа се поставят допълнителни файлове, които следват приетата конвенция за именуване плюс добавката „cover” ([град]-[институция]-М-[година с 4 цифри]-[код – инвентарен номер или др.]-[име на ръкописа или разпознаваема част от него]_[брояч]_cover.tiff) и съдържат корицата/кориците или заглавната страница.

При компресирането и конвертирането на TIFF файловете от основния дигитален архив в JPEG (или PDF) файлове за използване и предоставяне на достъп до тях в онлайн системи имената на папките и файловете се запазват като се променят само разширенията на файловете (.jpg или .pdf). Това дава възможност за добра оперативност при забелязани проблеми с онлайн достъпването на файлове с оглед на нови обработки и замени.

Получаването на имена по определена конвенция (схема) на дигитализирани книги е подобен на този за ръкописите:

- Име на папка: [град]-[институция]-[година с 4 цифри]-[код – инвентарен номер или др.]-В-[име на книгата или разпознаваема част от него].

- Име на файл: [град]-[институция]-[година с 4 цифри]-[код – инвентарен номер или др.]-В-[име на книгата или разпознаваема част от него]_[брояч].tiff

- [брояч]: 001, 002, 003..... (поредност на страници).

- В – book, сигнатура за визуално разпознаване на папки и файлове на дигитализирани книги.

Имената на файловете и папките се изписват задължително на латиница, като се спазва принципът за транслитерация от националната азбука по официално утвърдена таблица за конверсия. Не се допускат интервали или символи, различни от букви и цифри, с изключение на тире. Думите при необходимост се разделят със средно тире: „-“, долно тире („_“) се използва само при броячите в имената на файловете.

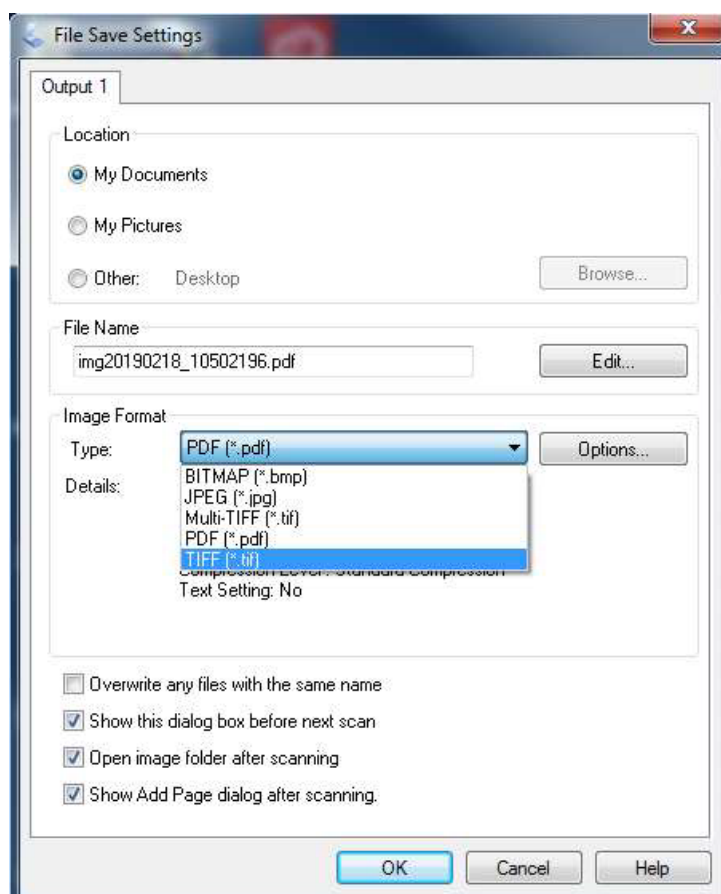
Всички TIFF (JPEG, PDF) файлове в папката на даден ръкопис или трябва да бъдат с еднакви имена, с изключение на брояча в края на името, който започва винаги от 001 (изключение правят документи над 1000 страници, при тях започва от 0001) и нараства последователно със стъпка 1 без пропуски.

Задължително при сканиране се проверяват файловете за наличие и поредност.

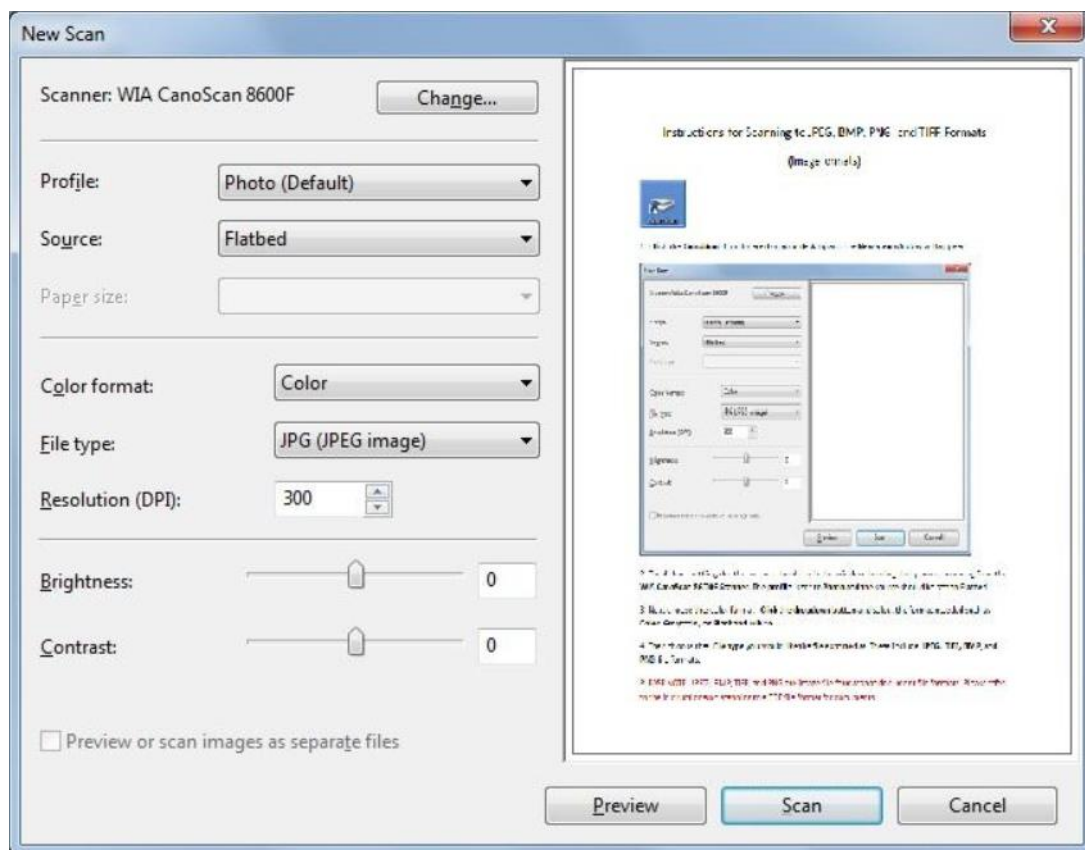
8. Настройване на техниката

Независимо от формата, който ще изберете при сканиране, обикновено скенерите и програмите, с които работите, имат възможност за запазване на сканираните файлове като TIFF, JPG и PDF. Следваните стъпки са:

1. Отидете на Файл > Запази като.
2. Изберете Файлов формат от падащото меню.
3. Изберете TIFF (JPEG, PDF или друг формат).



По подобен начин може да се променя цветовата характеристика и резолюцията при сканиране:



Изборът на характеристиките за запис на сканирания документ обикновено е от падащи менюта, но би могло да е налично и прецизно задаване на параметри.

9. Подходящи стандарти и спецификации за описание на цифровите обекти; изграждане на цифрови колекции/цифрови хранилища

9.1. Индексиране на цифровите обекти. Метаданни

Индексирането е много съществен етап от цифровизирането на обекти (славянски ръкописи и старопечатни книги). При цифровизацията на реални архиви, библиотеки, музейни експонати и др., може да се достигне до огромно количество цифрови образци и съответно без предварително индексиране (на всеки отделен образец по време на процеса цифровизация) може да се получат тривиални и досадни пречки за последващо обработване на тези образци. Друго възможно затруднение може да се получи при мрежово споделяне на ресурси и многоплатформеност, което се преодолява чрез използване на утвърдени стандарти за мета описание. При описанието на ресурсите е необходимо да се осигури съвместимост и функционалност на цифрова информация, да се създаде възможност за взаимодействие между различни хранилища на данни и системи. Освен това ресурсите се описват, така че да се подобри точността при търсене на информация и за тази цел описанието на ресурсите включва метаданни, описание на ниво колекция, контролни речници, уникални идентификатори.

Индексирането на цифрови ресурси се извършва посредством метаданни като се описва съдържанието на цифровия обект и кратко описание на файловете по определени признаци и класификации.

9.1.1. Стандарт Dublin Core

Dublin Core осигурява просто и стандартизирано множество от условия за описване на информацията онлайн по такъв начин, че да може да бъде лесно намирана. Dublin Core се използва широко за описване на дигитални материали – видео, звук, изображение, текст и съставни медии като уеб страници. Реализациите на Dublin Core обикновено се осъществяват с използване на XML (Extensible Markup Language) и са

Resource Description Framework базирани. Dublin Core е дефиниран от ISO Standard Z39.85-2001.]

Семантиката на Dublin Core е зададена и се поддържа от международна група от професионалисти от различни дисциплини – библиотекознание, компютърни науки, кодиране на текст и други, свързани области на науката и практиката.

Стандартът Dublin Core включва две нива: просто (simple) и квалифицирано (qualified). Първото ниво (Simple Dublin Core) се състои от 15 елемента; второто ниво (Qualified Dublin Core) включва три допълнителни елемента – читатели (Audience), произход (Provenance), притежател на правата (RightsHolder), както и група от елементи за уточняване, които разграничават семантиката на елементите, така че да могат да бъдат полезни при откриване на източници.

Simple Dublin Core Metadata Element Set (DCMES) се състои от следните 15 елемента:

- Заглавие (Title)
- Създател (Creator)
- Тема (Subject)
- Описание (Description)
- Издател (Publisher)
- Приносител (Contributor)
- Дата (Date)
- Тип (Type)
- Формат (Format)
- Идентификатор (Identifier)
- Източник (Source)
- Език (Language)
- Връзка (Relation)
- Обхват (Coverage)
- Права (Rights)

Заглавието представлява името, дадено на ресурса. Обикновено заглавието е името, под което ресурсът е формално известен.

Създателят е отговорен за създаването на ресурса. Примери за приносител включват човек, организация или учреждение.

Темата на ресурса може да бъде представена, като се използват ключови думи, ключови фрази или класификационни кодове. Добра практика е използване на контролен речник.

Описанието може да включва резюме, списък на съдържанието, графично представяне или свободен текст за ресурса.

Издателят е отговорен за предоставянето на ресурса в наличност. Примери за приносител включват човек, организация или учреждение.

Приносителят е отговорен за осигуряване на ресурса. Примери за приносител включват човек, организация или учреждение.

Датата е момент или период от време, свързан със събитие в жизнения цикъл на ресурса. Датата може да се използва за изразяване на временна информация на произволно ниво на подробност.

Типът представлява естеството или жанра на ресурса. Добра практика е използване на контролен речник като DCMI Type Vocabulary [DCMITYPE].

Форматът представлява файловия формат, физическо средство или мерки на ресурса. Примери за мерки включват размера и продължителността. Добра практика е използване на контролен речник като списъка на Internet Media Types [MIME].

Идентификаторът е недвусмислена референция към ресурса в даден контекст. Обикновено ресурсът се идентифицира чрез символен низ, съобразен с формалната система за идентифициране.

Източникът е ресурса, от който описаният ресурс е извлечен. Описаният източник може да бъде извлечен от свързан източник изцяло или отчасти. Добра практика е да се идентифицира свързания източник чрез символен низ, съобразен с формалната система за идентифициране.

Езикът на ресурса обикновено се определя с помощта на контролен речник.

Свързаният източник се идентифицира чрез символен низ, съобразен с формалната система за идентифициране.

Обхватът представлява пространствена или времева тема на ресурса, пространствената приложимост на ресурса. Пространствената тема може да

бъде именувано място или място, определено чрез неговите географски координати. Времевият период може да бъде период, дата или редица от дати.

Правата съдържат информация за правата над ресурса. Обикновено информацията за правата включва излагане на различни права за собственост, свързани с ресурса, включително и права за интелектуална собственост.

Всеки Dublin Core елемент е опционален и може да бъде повторен. Dublin Core Metadata Initiative (DCMI) задава стандартни начини за уточняване на елементи и съдейства за използването на кодиращи схеми и речници. В Dublin Core не са определени правила за наличието или използването на елементи.

Qualified Dublin Core съдържа допълнителни елементи, които ограничават и правят по-специфично значението на всеки елемент. Един уточнен (refined) елемент споделя значението на неквалифицирания елемент, но с по-ограничен обхват.

Освен елементи за уточняване **Qualified Dublin Core** включва множество от препоръчителни кодиращи схеми, проектирани с цел да подпомогнат интерпретирането на стойността на даден елемент. Тези схеми включват контролни речници и формални означения или правила. Една стойност, изразена чрез използване на кодираща схема може по този начин да бъде избрана от контролен речник или форматиран символен низ в съответствие с формално означение.

10. Характеристики за цифрова обработка на изображенията според последващи използване и достъп

Както бе отбелязано и в началото, през последните десетилетия много библиотеки по света работят по проекти за дигитализацията и онлайн каталогизацията на славянските ръкописи. Един от последните примери в тази насока е разработваната База данни по проект „Южнославянски кописи и скриптории от XIV век (палеографска атрибуция и онлайн реперториум)“ (<https://kopisti14.kmnc.bg/>). Използвали наличните данни от всички каталози (книжни и онлайн) и научни публикации за ръкописите, като по този начин е включена най-актуалната информация. Основна отличителна черта на онлайн базата данни е, че тя е отворена и има възможност за допълване. По този начин в бъдеще би могло да се извърши разширяване, както времево, така и пространствено, с прибавяне на нови данни за ръкописите. Стремелжът е към максимална стандартизация за възможно най-обективното описание, улесняващо използването и изследването на обектите. Базата с данни оперира паралелно на български и на английски език, а благодарение на стандартизацията в бъдеще би било лесно прибавянето и на други езици.

Базата данни на проект „Южнославянски кописи и скриптории от XIV век (палеографска атрибуция и онлайн реперториум)“ се отличава с това, че е онлайн и има предимства, които се дължат на дигиталното обработване на голям брой данни.

В рамките на проекта всеки писмен ресурс, още наричан обект „ръкопис“, се дигитализира и описва по дескриптивен (описателен) модел, базиран на наличните за него данни и знание. Описателният модел на обект „ръкопис“ включва базови дескриптори на нива: „Идентификация“, „Описание“ и „Палеография“. Трите основни нива на знание и стойностите на метаданните им са определени след задълбочен анализ и научно изследване на същността на представените обекти, както и на областта на книжовно наследство, към която принадлежат. Нивото „Идентификация“ съдържа общи идентифициращи аспекти на обект „ръкопис“ като наименование, произход, местонахождение, типология, датировка, сигнатура. Нивото „Описание“ покрива дескриптори, отнасящи се до ортографията, материала на основата (хартия, пергамент и др.), тялото на ръкописа, броя листове, размерите, библиографията и др. Нивото

„Палеография“ представя основни размери, декоративни елементи, типология и наклон (в градуси) на писмото, кописти (писачи) и др.:

❖ Ръкопис

➤ Идентификация

- Местонахождение:
Сигнатура:
- Част:
- Типология:
- Датировка (години)
От:
До:
Датировка - бележки:
- Произход - бележки:

➤ Описание

- Материал
Материал - бележки:
- Правопис
- Размери (mm)
Височина:
Ширина:
- Брой листове
Листове di guardia в начало:
Тяло на ръкопис:
Бележки:
- Библиография
- Източник:
Страници:
Препратки

➤ Палеография

- Типология на писмо:
Типология на писмо - бележки:
- Кописти
Копист:
Копист - бележки към ръкописа:
Лл. на ръкопис от този копист:
Колофон, в който се говори за копист
- Размер на писмо (mm)
Бележки:
Препратки

Изображения

По отношение на базата данни се планира разработка на софтуерни инструменти за подробно разглеждане и изследване на отделна буква и нейните елементи, откриване на сходство между двойки ръкописи и/или части от тях, изписването на букви, техните елементи; инструменти за подпомагане на съвместната изследователска работа на учените, и др. Целта е достигане и ефективно проучване на големи количества дигитализирани писмени ресурси с предоставяне на инструменти за по-висока адаптивност и ефективно взаимодействие между изследователя и дигиталния реперториум¹¹.

¹¹ Скарпа, М., Рипаранте, М., Панева-Маринова, Д., Гойнов, М., Лучев, Д. 2022. Онлайн база данни на южнославянски ръкописи от XIV в. Старобългаристика, XLVI (2022), 1, 181-202

11. Насоки в развитието на дигитализацията на славянски ръкописи.

OCR (Optical Character Recognition, оптично разпознаване на символи) за текст на печатен модерен стандартен език може да се счита за решена задача с процент грешки на знаците доста под 1%. Това обаче не е така за различните етапи на исторически текстове. Ранните печатни книги и в още по-голяма степен ръкописите представляват значителни затруднения поради хетерогенността, лошото качество понякога на изображението и липсата на стандартизация. Напоследък обаче извършването на OCR за ранни печатни книги и църковнославянски издания се подобрява значително. Все още обаче има много проблеми, когато става въпрос за сканиране на ръкописни текстове: традиционните OCR модели се провалят, когато се използват върху ръкописи, където резултатите са значително по-лоши, отколкото при печатен текст и практически неизползваеми в повечето случаи. Процесът, който се използва по отношение на ръкописите е HTR (handwritten text recognition, разпознаване на ръкописен текст).

В своя статия А. Рабус¹² разглежда възможностите за автоматично разпознаване на текст на модели на невронни мрежи, специално обучени да разпознават различни стилове на църковнославянски писмо в рамките на софтуерната платформа Transkribus. Реалната производителност показва, че специално обучените модели, като цяло, разпознават правилно простите (не надстрочни) знаци през повечето време. Процентът на грешки е по-висок при горни букви, съкращения и разделяне на думи. Комбинираните модели, състоящи се от данни за обучение от различни източници, са в състояние да транскрибират различни стилове на славянско писмо с нисък процент на грешки. Автоматичното разпознаване на текст с помощта на Transkribus и представените моделите в тази статия може да спомогнат за подобряване на ефективността на процеса на цифровизиране на славянските ръкописи.

През 2019 г. READ-COOP (<https://readcoop.eu/>) е създаден като част от модела за устойчивост на Transkribus, платформа за разпознаване на ръкописен текст (HTR), която изгражда значителна потребителска общност сред учени по хуманитарни науки и институции за културно наследство по

¹² Rabus, A. Manuscripts: a Neural-Network Approach Using Transkribus Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach Using Transkribus. Scripta & e-Scripta, 2019, 19, 9-32.

време на първоначалния си период, финансиран с безвъзмездна помощ от ЕС. В допълнение към предлагането на графичен потребителски интерфейс за транскрибиране на текстове, коригиране на транскрипции и обучение на модели за машинно обучение за изпълнение на HTR за нови езици и стилове на почерк, Transkribus предоставя инструменти за автоматично подравняване на съществуващите транскрипции с изображения на страници. Транскрипциите могат да бъдат дигитализирани (в смисъл да бъдат преобразувани в машинно четим текст, а не просто в сканирани изображения) с високо ниво на точност, като се използва достъпен в търговската мрежа софтуер за оптично разпознаване на символи (OCR), като ABBYY FineReader, и след това подравнени с изображенията на ръкописа с помощта на инструментите, предоставени от Transkribus. Това може значително да ускори процеса на разработване на данни за обучение за различни скриптове и почерци на писци. Подобен подход би могъл да се използва за подравняване на отделни ръкописни преписи от цифрови издания с фотографски факсимилета, където съществуват¹³.

В процес на развитие и тестване на модели спрямо славянските езици са и други системи за HTR, като eScriptorium. Платформата eScriptorium е изрично проектирана с насоченост към нелатинските скриптове и в момента е най-широко известна и използвана сред арабисти и хебраисти¹⁴.

В зависимост от това как текстът е бил аотиран и опциите за търсене, предоставени на потребителя, може да бъде възможно да се търси въз основа не само на поредици от букви (както може да се направи при обикновено търсене в пълен текст), но също и въз основа на граматически или семантични особености на текста.

Някои автори виждат развитието на цифровизацията на изследванията на средновековните славянски ръкописи като насочени към създаването в бъдеще на многократно използвана, оперативно съвместима инфраструктура (трансгранична и интердисциплинарна) – не като гигантска техническа платформа, а като екосистема от данни, модели, фрагменти от код, уроци. Това е екосистема, в която ръчно аотирани текстове се включват като данни, тези данни се преконфигурират като дървесни¹⁵

¹³ Dombrowski, Q. From Annotation to Modeling: Computational Horizons for Medieval Slavic Studies. Scripta & e-Scripta, 2021, 21, 11-21.

¹⁴ Thompson, T. Using Handwritten Text Recognition (HTR) Tools to Transcribe Historical Multilingual Lexica. Scripta & e-Scripta, 2021, 21, 217-231.

¹⁵ Дървовидна структура от данни.

банки, които се използват за обучение на NLP¹⁶ модели, които могат да се използват дори от хора с ограничени умения за кодиране.

В този контекст на дигитализация на средновековни славянски ръкописи се включва и изследователският проект „Южнославянски кописи и скриптории от XIV век (палеографска атрибуция и онлайн реперториум)“, който има за цел да създаде ИТ инструменти за изучаване и изследване на славянски ръкописи, почерци и кописи, сътворени през XIV в. Разработката предоставя средства за структуриране, обработка, управление, визуализиране и анализиране на цифровите аналози на тези ценни артефакти на книжовното наследство, като значително улеснява процесите на тяхното изследване и изучаване.

¹⁶ Natural language processing (NLP) – компютърната обработка на естествен (човешки) език е подобласт на науката за изкуствения интелект и компютърната лингвистика.

12. Техники на изкуствения интелект за разпознаване на ръкописен текст

Автоматичното разпознаване на ръкописен текст (Handwritten Text Recognition – HTR) в контекста на средновековното славянско писмено наследство представлява една от най-динамичните и високотехнологични области в дигиталните хуманитарни науки. За разлика от класическото оптично разпознаване на символи (виз. OCR), което е проектирано за стандартизиран печат, HTR системите разчитат на усъвършенствани архитектури на изкуствения интелект (AI) и машинното самообучение (Machine Learning), способни да моделират сложни вариации в индивидуалните изписвания, деградацията на носителите и палеографските особености на историческите шрифтове.

Архитектури на невронни мрежи за HTR

Съвременните софтуерни платформи за транскрипция (като Transkribus, Loghi или Kraken) и специализираните изследователски модели се базират на хибридни дълбоки невронни мрежи (Deep Neural Networks), комбиниращи два основни компонента:

- **Конволюционни невронни мрежи (Convolutional Neural Networks – CNN):** Те отговарят за компютърното зрение. CNN анализират дигиталното изображение на ниво пиксели, извличайки визуални характеристики, текстури, контури и геометрични форми на буквите, независимо от фоновия шум (петна, пробиви на мастило, гънки по пергаментата).
- **Рекурентни невронни мрежи (Recurrent Neural Networks – RNN), по-конкретно Long Short-Term Memory (LSTM) мрежи:** Тъй като ръкописният текст е последователност от знаци, LSTM мрежите анализират контекста. Те предсказват най-вероятния следващ символ или дума на базата на предходните, което е критично при слети текстове или полуизбледнели пасажии.
- **Трансформърни модели (Transformers):** Напоследък архитектури като Vision Transformers (ViT) започват да изместват LSTM, тъй като позволяват паралелно обработване на контекста и улавят дългосрочни зависимости в структурата на изреченията с много по-висока точност.

Предизвикателства пред славянската палеография

Обучението на AI модели за средновековна кирилица и глаголица е съпроводено със специфични палеографски и лингвистични предизвикателства, които изискват специализирани настройки на алгоритмите:

- **Непрекъснато писмо (Scriptura continua):** Липсата на интервали между думите в ранните славянски кодекси затруднява сегментирането. AI моделите трябва да се обучават да разпознават границите на думите не по разстоянието между тях, а чрез езикови вероятностни модели.
- **Лигатури и надредни знаци:** Използването на сложни лигатури (слети букви), титли (знаци за съкращаване на думи) и надредни букви изисква мрежата да разпознава символи на няколко нива над и под основната линия на писане.
- **Еволюция на шрифтовете:** Преходът от устав към полуустав и бързопис (курсив) променя радикално геометрията на буквите. Един модел, обучен на устав от XIV век, показва ниска точност при бързопис от XVII век, което налага създаването на профилирани или мултифондови модели.

Процес на обучение и концепцията за "Ground Truth"

Ефективността на изкуствения интелект зависи пряко от качеството на т.нар. Ground Truth (базов обучителен масив). Процесът протича през три основни стъпки:

- **Дигитално сегментиране:** Изкуственият интелект автоматично открива оформлението на страницата (Layout Analysis) – определя границите на текстовото поле, колоните, декоративните инициали и чертае базови линии (baselines) за всеки ред.
- **Експертна транскрипция:** Палеографи въвеждат ръчно точния дипломатически или критически текст за ограничен брой страници (обикновено между 50 и 100 страници за начално обучение).
- **Итеративно обучение (Training & Fine-tuning):** Невронната мрежа сравнява дигиталното изображение с предоставения текст хиляди пъти, докато процентът на грешка (Character Error Rate – CER) падне под приемливите 5–10%. Интегрирането на тези AI техники позволява мащабна обработка на огромни масиви от славянски ръкописи. Веднъж обучен, моделът може да транскрибира хиляди страници за минути, превръщайки статичните растерни изображения (TIFF/JPEG) в напълно

четим, структуриран и подлежащ на компютърен и лингвистичен анализ текстов формат.

Внедряването на техниките на изкуствения интелект и в частност на архитектурите за автоматично разпознаване на ръкописен текст (HTR) бележи фундаментален научно-технологичен обрат в изследването на средновековното славянско писмено наследство. Преходът от традиционните статични цифрови архиви към динамични, машинночетими текстови масиви отваря непознати досега хоризонти пред дигиталните хуманитарни науки. Чрез синергията между компютърното зрение (CNN) и контекстуалния езиков анализ (RNN/Transformers), съвременните алгоритми успешно преодоляват критични палеографски предизвикателства като липсата на интервали между думите (*scriptura continua*), избледнелите мастила и сложните системи от съкращения и надредни знаци.

Въпреки високата ефективност на изкуствения интелект, постигането на минимален процент на грешка (CER) остава пряко зависимо от експертната намеса на палеографи и филолози при подготовката на прецизни обучителни масиви (Ground Truth). Изкуственият интелект не замества учения, а функционира като мощен аналитичен инструмент, който елиминира рутинния труд по първичната транскрипция на източниците.

В перспектива, масовото прилагане на HTR системи върху дигитализираните колекции от славянски ръкописи ще позволи създаването на мащабни, взаимно свързани корпуси. Това не само ще улесни глобалния достъп до палеографските паметници, но и ще положи основата за провеждането на компютърно-лингвистични, статистически и текстологични изследвания в обеми, които до този момент бяха непосилни за традиционната хуманитаристика.

13. Заключение

Формулирането, систематизирането и стриктното научно-практическо прилагане на специализирани изисквания към цифровизацията на средновековни славянски ръкописи представлява основополагащ етап в съвременната стратегия за дългосрочно опазване, интердисциплинарно изследване и глобална социализация на това уникално културно-историческо наследство. Славянските кодекси притежават изключително специфична палеографска, кодикологична и физико-химична идентичност. Деликатният характер на пергаментовите и ранните хартиени носители, специфичният химичен състав на средновековните мастила и пигменти, наличието на микроскопични маргинални бележки, сложни орнаментални мотиви, миниатюри и водни знаци (филигранни) изискват методология, която излиза далеч извън рамките на стандартното репродуциране на документи. Процесът по тяхната дигитализация не може да бъде механичен акт; той е високотехнологична експертна дейност, балансираща между консервационната етика и софтуерно-хардуерните стандарти.

Успешната реализация на подобни проекти е възможна единствено чрез интегрирането на интердисциплинарен подход. Колаборацията между палеографи, лингвисти, реставратори, архивисти и ИТ инженери позволява дефинирането на технически параметри, които едновременно гарантират абсолютната физическа безопасност на артефактите и безкомпромисното качество на дигиталното съдържание. В този смисъл, въвеждането на задължителни критерии като използването на безконтактна планетарна апаратура за сканиране с V-образни люлки, контролираното студено осветление без UV и IR спектри, както и прецизното хроматично калибриране чрез профилиране на сензорите, съставляват критичния минимум за защита на оригиналите от механични и термични деформации по време на сесиите.

От технологична гледна точка, създаването на дигитални масиви от най-висок клас (Master Copies) изисква персистентно съхранение в некомпресирани формати без загуба на данни (като RAW и TIFF) с висока пространствена и дълбочинна разделителна способност. Това осигурява информационна изчерпателност на цифровия файл, позволяваща улавянето на най-фините детайли от текстурата на материала и наслояванията на мастилото. Успоредно с това, залагането на отворени стандарти за метаданни (като TEI XML за текстова транскрипция и Dublin Core/METS за библиографско описание) е гаранция срещу технологичното остаряване на

софтуерните среди и осигурява оперативна съвместимост между отделните академични бази данни.

В дългосрочен план, дигитализацията, подчинена на строги научни изисквания, рефлектира върху развитието на хуманитаристиката в няколко направления. От една страна, тя минимизира необходимостта от пряк физически контакт със славянските ръкописи, ограничавайки тяхното естествено амортизиране и деградация. От друга страна, научно издържаните дигитални реплики, интегрирани в модерни уеб-базирани платформи и протоколи (като IIF – International Image Interoperability Framework), премахват географските и институционалните бариери пред изследователите. Това осигурява равен достъп до източниците и отваря пътя за прилагане на иновативни методи от дигиталните хуманитарни науки – включително компютърен палеографски анализ, мултиспектрално изследване на заличени текстове (палимпсести) и разработване на специализирани OCR алгоритми за автоматично разпознаване на средновековна кирилица и глаголица. По този начин писменото наследство на средновековния славянски свят се превръща в жив, оперативен съвместим елемент от глобалната дигитална памет на човечеството.