УДК 02:004:778.19](477)

Депозитарна цифрова бібліотека — перспективний проект інформатизації Книжкової палати України

Галина Гуцол,
заступник директора з наукової роботи Книжкової палати України

У статті йдеться про новий перспективний проект створення електронних ресурсів Книжкової палати України як складової національного інформаційного фонду.

В статье говорится о новом, перспективном проекте разработки электронных ресурсов Книжной палаты Украины как составляющей национального информационного фонда.

The article says about the new, perspective project of working out of electronic resources of the Book Chamber of Ukraine as to a component of national information fund.

Ключові слова: *інформатизація, оцифрування, електронний ресурс, цифрова бібліотека, цифровий архів, Державний архів друку.*

Проблема інформатизації бібліотек та архівів, оцифрування документів і розроблення повнотекстових електронних ресурсів досить популярна в наш час і обговорювана в засобах масової інформації. Проте проекти інформатизації Книжкової палати України є настільки ж глобальними й унікальними, як і її фонд. Ці проекти завжди спрямовані на вирішення завдань у масштабах країни і призначені для задоволення потреб користувачів не лише в Україні, а й поза її межами.

Питання про створення електронного фонду видань Державного архіву друку Книжкової палати України і на його основі Депозитарної цифрової бібліотеки розглядалася в статті М. Сенченка "Депозитарна цифрова бібліотека — як єдина можливість вирішення проблеми книгосховищ" [1], але анонсово, як намітки до проекту, і не дає повного уявлення про цей проект. Пропонована стаття вперше знайомить читача з результатами розроблень щодо створення та використання електронного фонду Депозитарної цифрової бібліотеки, що і є свідченням її наукової новизни.

Наукові розробки здійснено у відповідності до теми наукових досліджень Книжкової палати України 2008 року "Розроблення й актуалізація електронних ресурсів сучасної та ретроспективної національної бібліографії України" (державний реєстраційний № 0108U004890).

Метою статті є висвітлення нових перспективних розроблень у галузі інформатизації Книжкової палати України, створення та використання нових електронних ресурсів для інформаційного забезпечення суспільства.

Ера цифрових технологій вимагає глобальних рішень щодо способів доступності великих документально-інформаційних масивів, таких, як бібліотеки й архіви, фонди яких традиційно перебувають на паперових носіях. За винятком абонементу в бібліотеках, вони доступні лише для локального використання в приміщеннях установ-фондоутримувачів і, відповідно, для обмеженого кола користувачів. Це ж стосується і фонду Державного архіву друку Книжкової палати України — головного сховища всіх видів видань, випущених в Україні з 1917 р. [2], який є ще більш закритим для використання, ніж фонди інших установ. Кількість видань у фонді станом на 29.09.2008 р. становить 13 292 242 одиниць зберігання, в тому числі:

- книг і брошур — 895 329;
- газетних видань — 10 063 310;
- періодичних і продовжуваних видань (крім газет) — 280 760;
- нотних видань — 18 530;
- образотворчих видань — 83 842;
- картографічних видань — 1 979;
- аркушевих текстових видань — 1 948 492.

Це унікальне зібрання творів друку, випущених видавництвами України, починаючи з 1917 р., яке перебуває під охороною держави та є її власністю [2]. Інтерес до фонду з кожним роком зростає пропорційно збільшенню кількості наукових досліджень щодо висвітлення історії держави і ролі в її творенні окремих особистостей.

Однак збільшення кількості звернень до фонду негативно впливає на стан видань, які під впливом часу втратили первісний вигляд і перебувають у критичному стані. Існує реальна загроза втрати унікальних документів, що становлять історичну і культурну цінність. Така ситуація обумовлена тим, що, на відміну від бібліотек, морально застарілі і фізично зношені фонди яких підлягають списанню, Книжкова палата України довічно зберігає в одному єдиному примірнику видання, які надходять до неї згідно із Законом України "Про обов'язковий примірник документів" [3].

Найбільшу цінність фонду Державного архіву друку становлять видання 1917—1940 рр., а також фонд спеціального зберігання, що складається з видань 1917—1976 рр., заборонених органами цензури. Значна кількість видань цього періоду представлена в одному, унікальному примірнику, що зберігся в Україні. Саме з метою якнайкращого збереження видань Книжкова палата України надає доступ до фонду лише науковцям і вимагає при цьому лист від провідної установи із зазначенням теми дослідження та довідки від бібліотек про відсутність затребуваних видань в їхніх фондах.

Проте навіть таке суворе обмеження на використання фонду не може запобігти згубному впливу на його стан. Адже лише одна операція ксерокопіювання "зістарює" видання на сім (!!!) років. Єдиним правильним вирішенням цієї проблеми може стати збереження копії видань на інших носіях інформації, найсучаснішими з яких є електронні носії. Це дасть можливість одночасно розкрити фонди, забезпечивши доступ до них для необмеженої кількості користувачів з усього світу за допомогою Інтернету і зберегти оригінали видань для майбутніх поколінь.

Першочергово необхідно перевести на електронні носії найціннішу частину фонду — видання 1917—1940 рр. та спецфонд. Проект щодо їхнього оцифрування отримав назву "Переведення на електронні носії історичних фондів Державного архіву друку Державної наукової установи "Книжкова палата України імені Івана Федорова"" і був схвалений Колегією Держкомтелерадіо України на засіданні 24 квітня 2008 р. У її рішенні, зокрема, сказано: "Підтримати пропози-

цію Державної наукової установи "Книжкова палата України імені Івана Федорова щодо проведення робіт з перенесення на електронні носії видань фонду Державного архіву друку" [4].

Наслідком цього стало розроблення технічного завдання до проекту, продовженням якого має стати галузева цільова програма.

Розробленню проекту передувала велика підготовча робота, яка складалася з: вивчення літератури за темою та світового досвіду у справі. З цією метою в Книжкову палату України для проведення семінару з оцифрування фондів було запрошено Сьюзен Бенц, співробітницю Бруклінської публічної бібліотеки, США, яка проходила наукове стажування в Україні за Програмою Фулбрайта. Пані Сьюзен була керівником проекту з оцифрування фондів у своїй бібліотеці і її практичний досвід з планування [5] і реалізації проектів виявився неоціненним для фахівців Книжкової палати України.

Після підготовчої роботи було окреслено коло питань, на які проект мав дати відповідь:

1. Якої мети ми хочемо досягти?
2. Що і в якому вигляді хочемо отримати?
3. Які прогнозні обсяги інформації отримаємо?
4. Як використовуватимемо результати?
5. Які технічні засоби необхідні для реалізації проекту?

Виходячи з цього, визначено основні завдання проекту:

1. Найбільш цінна з історичної точки зору і затребувана науковцями частина фонду Державного архіву друку Книжкової палати України, яка потребує першочергового переведення на електронні носії, містить:

- видання 1917—1940 рр., зокрема книжкові 104522 друк. од., газетні 1133010 друк. од. та журнальні 6376 друк. од.;
- видання 1917—1976 рр., заборонені органами цензури (спецфонд) — 17236 друк. од.

У найгіршому стані знаходяться газетні видання, надруковані на тонкому кислотному папері, який швидко руйнується і на деяких виданнях вже помітні втрати. Крім того, газетний текст зникає внаслідок випаровування фарби. Таким чином, якщо не вжити термінових заходів щодо виготовлення копії видань навіть за умови проведення реставраційних робіт, які зміцнюють і відновлюють структуру документа, він все одно втрачає цінність, бо стає нечитабельним.

Журнальні видання збереглися краще, тому що у свій час були скомплектовані у річні підшивки, скріплені і вставлені в палітурку. Завдяки обкладинці чи палітурці збереглися і книжкові видання.

Проте запити на видання почастішали, що вимагає ксерокопіювання, яке, як вже було зазначено вище, значно скорочує їхнє життя.

Отже: за допомогою оцифрування видань можливе одночасне вирішення двох протилежних завдань — збереження фондів і розкриття їх для необмеженого загального користування.

2. Існують різні підходи до оцифрування видань, що обумовлює параметри їхнього сканування і формати збереження інформації. Статистика відвідувань бібліотек свідчить, що найчастіше до них звертаються читачі з метою виконання навчальних завдань або написання наукових робіт. Виходячи з цього, метою оцифрування бібліотечних фондів (за винятком цінних і рідкісних) є якнайширше розкриття змісту видання, а не збереження його вигляду. Тому бібліотеки або сканують зображення сторінок з мінімальними параметрами — невисока роздільна здатність та налаштування кольору B&W (чорно-білий) чи Grayscale (відтінки сірого) і зберігають у графічних форматах з максимальним стисненням, або "розпізнають" зображення і зберігають у текстових форматах. Перевагою таких цифрових копій є невеликі обсяги отриманих файлів і, відповідно, зменшення вимог до апаратних і програмних засобів для їхнього збереження.

Книжкова палата України — Державний архів друку — має за мету, насамперед, зберегти цифрову копію видання, максимально відтворивши у зображенні його зовнішній вигляд як артефакту. Встановлено, що найкращого результату буде досягнуто за умови сканування повнокольорового зображення з роздільною здатністю 300 dpi та збереження кожної сторінки в окремому файлі в графічному форматі TIFF як такому, що містить максимальну інформацію про зображення.

Крім того, необхідно створити файл у форматі PDF, який відповідатиме одному виданню і міститиме зображення сторінок у послідовності їхнього розташування в друкованому оригіналі. Проте файли формату TIFF мають великі розміри (орієнтовно 5 Мб), тому для публікації в Інтернеті необхідне збереження зображень сторінок в іншому форматі. Загальноприйнятим для цього є формат JPEG, хоча останнім часом досвідчені фахівці в галузі Інтернет-технологій віддають перевагу економічному формату PNG.

Наприкінці 2007 р. Книжковою палатою України у співробітництві з Євро-Азійською Міжнародною Торговою палатою (ЄАМТП) здійснено пілотний проект з пробного оцифрування видань фонду Державного архіву друку — по одному примірнику книги, підшивки газет і журналів. ЄАМТП є неполітичною неприбутковою організацією з центральним офісом у м. Празі, Чеська Республіка, об'єднує компанії з різних країн світу і здійснює впровадження інноваційних технологій у галузі спеціальної техніки та залучення інвестиційних проектів з фондів Євросоюзу.

Виконання робіт з оцифрування було покладено на компанію SIA "Infodisk Media", асоційованого члена ЄАМТП. Вона має величезний досвід з оцифрування фондів бібліотек та архівів у різних країнах світу, бере участь у програмі OPEN ARCHIVE у країнах Скандинавії та Балтії, виконала ряд проектів з оцифрування найбільших державних архівів, у тому числі й Королівського архіву Швеції.

У результаті отримано високоякісні зображення сторінок видань, включаючи обкладинки і форзаци, а також електронний документ у форматі PDF, який є повною копією друкованого видання. Для визначення обсягів сканування та розмірів інформації для збереження в 2008 р. проведено моніторинг вищеназваного фонду на предмет визначення форматів видань і підрахунку кількості сторінок.

Моніторинг фонду книжкових видань здійснювався за даними карткового каталогу. Зважаючи на те, що в каталозі відсутні дані про обкладинки та форзаци (наявні приблизно в половині видань), до підрахованої кількості було додано по шість сторінок на кожне видання. Моніторинг журнальних і газетних видань проводився безпосереднім підрахунком сторінок у фонді, тому ці дані є точнішими. Результати моніторингу наведено в таблиці:

Ч.ч.	Вид видань	Кількість видань, друк. од.	Кількість сторінок	Формат
1.	Спецфонд 1917—1976 рр.	17236	1759384	A4
2.	Книжкові видання 1917—1940 рр.	104522	7718870	A4
3.	Журнальні видання 1917—1940 рр.	6376	1495392	A4
			640882	A3
4.	Газетні видання 1917—1940 рр.	1133010 (17075 річних підшивок)	453204	A1
			4078836	A3
	Всього:	**1261144**	**16146568**	

Виходячи з цього, підраховано обсяги електронного простору, необхідного для збереження цифрового архіву — близько 17,5 млн файлів загальним обсягом 100 Тб.

3. Картковий каталог Книжкової палати України існує з 1917 року, електронна база даних каталогу, яка називається "Книги за роки незалежності", містить відомості про видання з 1991 р. Одночасно з її створенням Книжковою палатою України проводилися наукові дослідження з ретроспективного розроблення фонду Державного архіву друку. Результатом є електронні бази даних бібліографічної інформації:

- газетних видань 1917—1937 рр.;
- журнальних видань 1917—1931 рр.;
- спецфонду 1917—1925 рр.;
- довідково-бібліографічних видань 1917—1922 рр.;
- бібліографічних і бібліографознавчих видань 1917—1921 рр.

У 2008 р. розпочато створення бази даних каталогу 1917—1990 рр. Метою цієї роботи є розроблення електронного каталогу всього фонду Державного архіву друку, який планується розмістити для загального доступу на веб-сайті Книжкової палати України.

У результаті реалізації проекту з оцифрування фонду буде отримано електронний архів, пошуковою системою до якого стане електронний каталог.

4. Виходячи з прогнозних обсягів електронного архіву, для організації збереження величезної кількості взаємопов'язаних графічних файлів і бібліографічних баз даних, необхідно дуже уважно підійти до вибору комплексу комп'ютерного обладнання і програмного забезпечення — системи збереження даних (СЗД).

Збереження архівів відрізняється від звичайного розміщення даних на дисках (дискових масивах) своїми унікальними вимогами. Ключовими з них є:

Висока масштабованість. Оскільки СЗД цифрового архіву стає основним місцем збереження даних, які поступово нарощуються упродовж тривалого часу, то вона повинна забезпечувати можливість нарощення електронного простору для збереження інформації (масштабованість) від одиниць до десятків, сотень або, навіть, тисяч терабайт. Така СЗД економічно ефективніша за системи зі сталим електронним простором, оскільки дає змогу планувати витрачання коштів на її нарощення відповідно до темпів нарощення обсягів інформації.

Незалежність від типу носія даних. СЗД цифрових архівів призначені для довготривалого зберігання даних і бажано, щоб вони зчитувалися у стандартному форматі, що не залежить від типу носія, і легко, без втрат, могли б бути перенесені на нове покоління носіїв інформації, які з'являються на ринку комп'ютерної техніки.

Показовим у цьому плані є випадок, який увійшов в історію, як "Ефект NASA" — американської космічної агенції, в якій назавжди втрачені архіви багаторічних подорожей на Місяць, тому що дані були записані на магнітних стрічках в унікальному форматі. Серед нового покоління пристроїв зчитування не знайшлося такого, який зміг би розпізнати цей формат.

Можливість архівування незатребуваної інформації. Цифровий архів призначений для накопичення і зберігання даних незалежно від того, як часто вони затребувані користувачами. СЗД цифрового архіву має забезпечувати економне використання електронного простору для збереження інформації за допомогою багаторівневого програмного архівування даних і відповідного розміщення файлових архівів. Для цього система повинна контролювати дату останнього звернення до файлу. Чим довше файл не був затребуваний, тим щільніше він має бути заархівований і розміщений на нижчому рівні. При цьому система має забезпечити швидке здобуття файлу в будь-який час і онлайновий доступ до нього.

Універсальна динамічна підтримка додатків. У сучасних СЗД дані, що належить архівувати, генеруються множиною різноманітних програмних засобів (додатків). Тому, щоб уникнути створення окремих "островів" архівування, що відповідають окремим додаткам, сучасна СЗД повинна пропонувати для всіх додатків абстрактні образи (на кшталт перехідних форматів, конверторів), що об'єднують всі дані в єдиний архів.

Гарантія незмінності даних в архіві. Так само, як друковане видання залишається незмінним упродовж всього терміну зберігання, так і його електронна копія має залишатися незмінною незалежно від терміну зберігання і частоти користування нею. Забезпечення незмінності даних у процесі довготривалого зберігання — одна з основних вимог до сучасних СЗД

Розглянувши всі вимоги Книжкової палати України до СЗД цифрового архіву, для реалізації завдань проекту компанією "Інком" був запропонований потужний відмовостійкий програмно-технічний комплекс формування цифрового архіву видань на основі двох серверів PrimeServer Lan2900R (активного та резервного), які забезпечують роботу з пристроєм цифрового архіву Centera у складі чотирьох архівних стійок Centera ємністю по 98 ТБ кожна для зберігання запланованих обсягів інформації. Сервери PrimeServer Lan2900R забезпечують також зовнішній Інтернет-доступ до архіву за допомогою бібліографічних баз даних, що мають бути ретрансльовані з використовуваного нині бібліографічного програмного забезпечення ProCite у MARK-SQL.

EMC являє собою спеціалізовану СЗД для файлових архівів, уперше запропоновану компанією EMC у 2002 році як систему збереження даних нового типу, спеціально призначену для он-лайн цифрових архівів. Centera використовує модель програмного забезпечення розподілених об'єктів, відому як Content Addressed Storage (CAS) – СЗД, що адресується за вмістом. CAS не використовує традиційних файлових систем і не потребує використання визначених носіїв даних, цій моделі не потрібна інтеграція на рівні ядра для серверних додатків. Примножуючий сінергетичний ефект усіх цих особливостей дав змогу створити повністю відмінну від попередніх архітектуру цифрового он-лайн архіву, що й уможливило виділення цієї системи в лінійку систем нового типу. СЗД Centera вже використовуються в Україні великими промисловими підприємствами, корпораціями, які працюють у галузі інформаційних технологій, телекомунікаційними компаніями тощо.

Отже, після відповіді на всі організаційні питання проекту необхідно розрахувати етапи його реалізації. Експеримент, здійснений співробітниками Книжкової палати України, довів, що проведення робіт тільки зі сканування книг спецфонду (17236 книг обсягом 1759384 сторінок) без корегування зображень у графічному редакторі за умови виконання їх двома особами, затягнеться на довгі 14 років. Обсяги сканування періодичних видань набагато більші, а їхні формати (А1) вимагають використання особливого обладнання, вартість якого в десятки разів перевищує вартість такого ж обладнання для сканування менших форматів (А4, А3).

Зважаючи на те, що видання швидко руйнуються (кришиться папір, зникає шрифт), переведення на електронні носії фонду, особливо газетних видань, необхідно провести якнайшвидше, залучивши до виконання робіт організації зі сторони, які мають значний досвід у проведенні таких робіт, відповідне обладнання і спеціалістів.

За даними компанії SIA "Infodisk Media", що здійснювала пробне оцифрування видань з фонду Книжкової палати України, виробнича потужність використовуваних нею сканерів за зміну (8 годин) становить:

— формат А1 — 500 сторінок;
— формат А2 — 700 сторінок;
— формат А3 — 800 сторінок;
— формат А4 — 1000 сторінок.

Виходячи з розрахунку використання десяти сканерів у дві зміни п'ять днів на тиждень, роботи з оцифрування книг, журналів і газет 1917—1940 рр. та спецфонду планується здійснити в чотири етапи, упродовж 2009—2012 рр.

Очікуваними результатами реалізації проекту є:

Перший етап — 2009 р.:

- програмно-технічний комплекс формування цифрового архіву фонду Державного архіву друку;
- цифровий архів газетних видань 1917—1938 рр.;
- Web-портал "Історичні фонди Державного архіву друку Книжкової палати України" з доступом через мережу Інтернет до електронного фонду газетних видань 1917—1938 рр.

Другий етап — 2010 р.:

- програмно-технічний комплекс формування цифрового архіву фонду Державного архіву друку з масштабованим вдвічі дисковим простором для зберігання;
- цифровий архів газетних видань 1939—1940 рр., спецфонду 1917—1976 рр. та журнального фонду 1917—1940 рр.;
- Web-портал "Історичні фонди Державного архіву друку Книжкової палати України" з доступом через мережу Інтернет до електронного фонду періодичних видань (газет і журналів) 1917—1940 рр. та книжкових видань 1917—1976 рр., заборонених цензурою.

Третій етап — 2011 р.:

- програмно-технічний комплекс формування цифрового архіву фонду Державного архіву друку з масштабованим утричі дисковим простором для зберігання;
- цифровий архів книжкових видань 1922—1931 рр. (книги 1917—1921 рр. відносяться до спецфонду);
- Web-портал "Історичні фонди Державного архіву друку Книжкової палати України" з доступом через мережу Інтернет до електронного фонду періодичних видань (газет і журналів) 1917—1940 рр., книжкових видань 1922—1931 рр. та видань 1917—1976 рр., заборонених цензурою.

Четвертий етап — 2012 р.:

- програмно-технічний комплекс формування цифрового архіву фонду Державного архіву друку з масштабованим в чотири рази дисковим простором для зберігання;
- цифровий архів книжкових видань 1932—1940 рр.;
- Web-портал "Історичні фонди Державного архіву друку Книжкової палати України" з доступом через мережу Інтернет до електронного фонду видань (книг, газет і журналів) 1917—1940 рр. та книжкових видань 1917—1976 рр., заборонених цензурою.

Доступ до оцифрованих видань, які є об'єктом майнового права, буде здійснено з дотриманням чинного законодавства про авторське право.

Висновки:

Реалізація проекту з переведення на електронні носії документів Державного архіву друку дасть змогу створити Депозитарну цифрову бібліотеку видань України і матиме значний соціальний ефект:

- забезпечить збереження на століття унікального історичного фонду видань України;
- розкриє фонди Державного архіву друку і введе їх до широкого наукового обігу;
- створить максимально повну, зручну в користуванні систему взаємопов'язаної багатоаспектної довідкової інформації про склад та зміст фонду;
- забезпечить оперативність у наданні повнотекстової інформації через відповідні комунікаційні засоби.

Вже після перших чотирьох років його реалізації, 2009—2012 рр., національний інформаційний фонд поповниться великим електронним інформаційним ресурсом — цифровим архівом газетних і журнальних видань 1917—1940 рр., книжкових видань 1922—1931 рр. та видань 1917—1976 рр., заборонених цензурою.

Проект здійснюватиметься з метою наповнення інформаційного простору держави й забезпечення бібліотек, навчальних закладів, наукових установ та всіх зацікавлених споживачів інформації з усього світу повнотекстовою електронною інформацією про видання України.

Перспективами проекту є створення державного електронного реєстру оцифрованих фондів бібліотек та архівів.

Список використаної літератури

1. *Сенченко М. І.* Депозитарна цифрова бібліотека — як єдина можливість вирішення проблеми книгосховищ / Микола Сенченко // Вісн. Кн. палати. — 2008. — № 2. — С. 3.
2. *Про* видавничу справу : Закон України [від 5 черв. 1997 р. № 318/97-ВР] // Урядовий кур'єр. — 1997. — 19 лип.
3. *Про* обов'язковий примірник документів : Закон України [від 9 квіт. 1999 р. № 595-XIV] // Урядовий кур'єр. — 1999. — 13 трав.
4. *Про* реалізацію рішення колегії Держкомтелерадіо ["Про необхідність перенесення на електронні носії фондів Державного архіву друку Державної наукової установи "Книжкова палата України імені Івана Федорова"] від 24 квітня 2008 року № 5/14 : наказ Державного комітету телебачення та радіомовлення України від 12 трав. 2008 р. № 120.
5. *Бенц С.* Огляд питань планування процесу переведення документів у цифрову форму / Сьюзен Бенц // Вісн. Кн. палати. — 2008. — № 2. — С. 13—15.
6. *Быков А.* Архивирование — важнейшая часть интеллектуальной системы хранения данных / Алексей Быков // Корпоративные системы. — 2007. — № 6. — С. 69—74.