УДК 930.25:004

О. Т. ФІТКУЛІН*

ВИБІР ТЕХНОЛОГІЙ ТА ПРОГРАМНИХ ЗАСОБІВ ДЛЯ СТВОРЕННЯ АРХІВНИХ КОПІЙ ВЕБ-САЙТІВ

Розглядається досвід деяких країн щодо зберігання веб-ресурсів. Аналізуються програмні засоби для створення архівних копій веб-сайтів.

Ключові слова: веб-сайт; архівне зберігання веб-сайтів; веб-краулер.

Центральний державний електронний архів України (далі – ЦДЕА України) як державний орган виконує завдання та функції держави з управління архівною справою та діловодством, забезпечує облік, збереженість електронних документів Національного архівного фонду та електронних інформаційних ресурсів і використання їх інформації[1].

Одним з важливих напрямків діяльності ЦДЕА України є створення архівних колекцій веб-сайтів у межах ініціативного документування. Їх формування спрямоване на зберігання культурної спадщини України для майбутніх поколінь. Для цього використовується веб-краулер**, який імітує веб-браузер і звертається за визначеними посиланнями відповідних HTML-сторінок, копіює їх, сканує наявність гіперпосилань на складові її змісту, супровідні файли та наступні HTML-сторінки і переходить за гіперпосиланнями, повторюючи свої дії, до наступних складових веб-сайту.

Під час створення архівних колекцій веб-сайтів особлива увага приділяється повноті завантаженої інформації. Контент***, який не був завантажений програмними засобами, довантажується вручну, і той, що не має відношення до тематики колекції, вилучається так само. При такому підході до зберігання електронної інформаційної спадщини України необхідні технології та засоби завантаження веб-сайтів, що забезпечують зберігання контенту з можливістю автоматизованої фільтрації джерел, моніторингу та контролю процесу копіювання.

Розуміючи цінність певної частини інформації, що розміщується в мережі Інтернет, у багатьох країнах розпочато дослідження щодо вирішення питання, які ресурси мережі Інтернет слід зберігати і за допомогою яких засобів це можна реалізовувати.

---

* *Фіткулін Олександр Тагірович* – головний спеціаліст відділу інформаційних технологій Центрального державного електронного архіву України.

** Веб-краулер – програма, що призначена для пошуку веб-сторінок у мережі Інтернет з метою подальшого їх зберігання.

*** Контент – узагальнюючий термін, що характеризує будь-яку інформацію, яка міститься на сторінках веб-сайту (зображення, аудіо, відео тощо).

Аналізуючи досвід різних країн, можна розподілити методики зберігання веб-сайтів на три групи:

1) країни, у яких національні закони про обов'язковий примірник* було розповсюджено на цифрові матеріали. Наприклад, Литва (з 1996 року зберігається обов'язковий примірник електронних документів, які мають велике значення для країни, а в 2002 року – для зберігання мережевих матеріалів було створено Архів електронних ресурсів як підсистему LIBIS (Литовська інтегрована бібліотечна інформаційна система)[2]. У Новій Зеландії обов'язковий примірник був розповсюджений на всі цифрові матеріали, у тому числі Інтернет-ресурси відкритого доступу, включаючи блоги**, вікі*** тощо, та створено Національний архів цифрового спадку для забезпечення постійного зберігання цифрової інформації про Нову Зеландію[3]. У Франції з 2006 року в автоматизованому режимі зберігаються всі мережеві матеріали національного домену****, а якщо їх автоматичне зберігання неможливе, ці матеріали вимагаються у видавництв[4];

2) країни, де зберігаються всі веб-сайти національного домену. Наприклад, Португалією скопійовано більше, ніж 130 мільйонів вебсторінок та надається вільний доступ до них[5]. В Об'єднаному Королівстві Британії з квітня 2013 року з національного домену UK завантажено 36 ТБ інформації[6]. У Фінляндії для копіювання, реєстрації та постійного зберігання Інтернет-публікацій створено архівну систему EVA[7];

3) країни, де вибірково зберігаються веб-сайти, що мають національне значення. Наприклад, в Австралії створено архів PANDORA[8], а для автоматизованого копіювання та описування веб-ресурсів, а також організації доступу до них було розроблено спеціальне програмне забезпечення PANDORAS. У Китаю з 2003 року реалізується проект щодо відбору, збереження, індексації та публікації веб-сайтів (веб-інформації) WICP (Web Information Collection and Preservation)[9], що пов'язані з історичними подіями, які мають велике значення для країни. У Нідерландах з 2006 року зберігаються спеціально відібрані мережеві ресурси на основі договорів з видавцями. Станом на початок 2010 року було скопійовано приблизно 2500 веб-сайтів[10].

---

* Обов'язковий примірник – екземпляр публікації, що в обов'язковому порядку, визначеному законодавством країни, надається бібліотекам та іншим установам з метою реєстрації і обліку, створення державних архівів тощо.

** **Блог – веб-сайт, основний зміст якого – записи, які регулярно додаються** і містять текст, зображення або мультимедіа.

*** **Вікі – веб-сайт, що дозволяє користувачам самостійно змінювати вміст** сторінок через браузер, використовуючи спрощену і зручнішу розмітку тексту.

**** **Домен**, або доменне ім'я – символьне ім'я, що допомагає знаходити адреси Інтернет-серверів.

У країнах, де закон про обов'язковий примірник розповсюджено на цифрові матеріали, зберіганням веб-сайтів займаються бібліотеки або створюються спеціалізовані бібліотечні архіви. Необхідно розуміти, що бібліотеки приймають для зберігання публікації, що тиражуються масово, а архіви займаються унікальними документами.

Копіювання веб-сайтів – це складний автоматизований процес, який виконується програмними комплексами, що відбувається шляхом індексування і збереження даних відповідно до попередньо встановлених параметрів.

Якість і повнота результатів завантаження залежить від програм, що використовуються та постійно вдосконалюються. Результат роботи веб-краулера – статичне представлення веб-сайту, незалежно від того, яким він був до копіювання. Статичними називаються веб-сайти, інформація в яких зафіксована у формі HTML-сторінок*, без використання додаткових програмних засобів для їх зміни за запитами користувачів. При кожному звертанні до веб-сайту для користувачів відтворюються відповідні HTML-сторінки. Щоб оновити інформацію на подібних сторінках, необхідно вручну внести зміни безпосередньо в HTML-код сторінки. На відміну від статичних, динамічні веб-сайти більш гнучкі в керуванні. Для роботи динамічних веб-сайтів використовуються різні технології, що дозволяють "конструювати" веб-сторінки "на льоту" на запити їх користувачів, а потім відтворювати їх. Динамічні веб-сайти можуть "підлаштовуватися" під своїх відвідувачів, реагуючи на їхні дії. Для цього використовуються технології серверних, клієнтських скриптів**, за допомогою яких і створюються сценарії роботи веб-сайтів при певних діях користувачів. Наприклад, сторінка у соціальній мережі, де відображено інформацію, пов'язану з користувачем, який увів свої ідентифікаційні дані.

Пропонуємо визначити такі критерії вибору програмних засобів для копіювання веб-сайтів:

– повнота інформації, що завантажується;

– багаторівневість копіювання інформації;

– можливість довантаження інформації з місця зупинки завантаження;

– можливість налаштування параметрів завантаження;

– відкритість програмного коду;

– безпека.

Орієнтація на ці критерії – дуже важлива умова під час вибору програмного засобу копіювання веб-сайтів, адже саме від нього залежить

---

* HTML-сторінка – веб-сторінка, написана мовою веб-розмітки HTML.

** Скрипт – мова програмування високого рівня для написання сценаріїв – коротких описів дій, які виконує система.

ефективність створення архівних колекцій веб-сайтів, зокрема, в ЦДЕА України. Розглянемо визначені критерії більш детально:

Повнота інформації, що завантажується. Критерій, викликаний необхідністю завантаження всієї інформації із вказаного веб-сайту, тобто всіх файлів, що знаходяться на ньому, незалежно від його складності. Разом з тим, часто файли зображень, відео, аудіо закриті для завантаження власниками веб-сайтів з метою захисту контенту. Такі файли доводиться завантажувати вручну.

Багаторівневість копіювання інформації. Дуже важливо, щоб у програмі копіювання веб-сайтів була можливість вибору рівнів завантаження, тобто в архівіста була можливість контролю меж переходу програми за посиланнями як всередині одного домену, так і на зовнішні ресурси відносно до нього. Оскільки для створення тематичних колекцій веб-сайтів необхідно копіювати лише інформацію, що тематично належить до створюваних архівних колекцій.

Можливість довантаження з місця зупинки завантаження інформації. Утрати Інтернет-з'єднання з будь-яких причин під час завантаження веб-сайту можуть призвести до припинення копіювання веб-сайту. Тому необхідно, щоб веб-краулер мав змогу продовжити копіювання веб-сайту з того місця, на якому сталася зупинка завантаження інформації.

Можливість налаштування параметрів завантаження. Веб-краулер повинен забезпечувати вибір таких параметрів завантаження веб-сайту: кількість одночасних з'єднань та кількість повторів автоматичного з'єднання, якщо відбулася втрата Інтернет-з'єднання. Необхідність наявності такого параметру, як кількість одночасних з'єднань, зумовлена тим, що деякі веб-сайти мають захист від перенавантаження каналу. Для цього встановлюється ліміт на загальну кількість одночасних звернень користувачів до контенту веб-сайту. При перевищенні цього ліміту, всі запити від джерела, що викликало перенавантаження, на практиці блокуються. Тобто, у випадку з веб-краулером, копіювання веб-сайту припиняється на період блокування. Це збільшує час, необхідний для повного копіювання веб-сайту.

Відкритість програмного коду. Важливість відкритого програмного коду полягає у незалежності від розробника програмного засобу, що застосовується для копіювання веб-сайтів. Удосконалення програми із закритим кодом та внесення змін може здійснювати лише розробник. У випадку завершення підтримки програмного продукту, з часом він стає застарілим та неактуальним для використання. На відміну від програм із закритим кодом, програми з відкритим кодом можуть бути модифіковані або допрацьовані, вільно розповсюджуватися, адаптовані для власних потреб. Завдяки цьому, такі програми частіше оновлюються, їх помилки швидше виявляються та виправляються.

Блокування зовнішніх посилань. Оскільки на веб-сайті можуть бути розміщені посилання на веб-ресурси, що не є необхідними для завантаження; виходячи з тематики створюваної колекції веб-сайтів, потрібно, щоб веб-краулер блокував можливість переходу за такими посиланнями у створених копіях. Важливим чинником є те, що зовнішні посилання повинні бути ізольовані, а не видалені взагалі. Тобто при переході за ними виводилось повідомлення про те, що ці посилання ведуть до веб-ресурсів, які не є частиною скопійованого веб-сайту.

Безпека. Необхідно, щоб програмний засіб був перевірений на безпечність і виконував лише ті фунції, які повинен виконувати. Досягти цього можна створенням комплексної системи захисту інформації в установі в порядку, визначеному законодавством України.

Розглянемо можливості програмних засобів, апробовані на практиці у ЦДЕА України:

HTTrack – безкоштовна програма з відкритим програмним кодом, розроблена Ксав'єром Роше і ліцензована відповідно до GNU General Public License Version 3 (ліцензія безкоштовного програмного забезпечення). Програма має багато параметрів для налаштування завантаження інформації: кількості одночасних з'єднань, їх максимальної швидкості, кількості повторів, рівнів копіювання інформації, налаштування проксі-сервера, можливості блокування зовнішніх посилань тощо.

HTTrack може оновлювати попередньо скопійовані веб-сайти та відновлювати завантаження у випадку розриву з'єднання з мережею Інтернет. Має базову версію командного рядка та три версії із графічним інтерфейсом (WinHTTrack, WebHTTrack та HTTraQt).

Недоліками програми є те, що вона не завантажує деякі файли (зображення, аудіо та відео, css* та інші), якщо посилання на них сформовані за допомогою мови програмування Java-script. Посилання, сформовані таким чином, є динамічними, тобто формуються або виконують ті чи інші дії лише у відповідь на дії користувача, а жодна з програм для копіювання веб-сайтів на сьогодні не може їх імітувати. Тому недовантажені файли необхідно зберігати вручну. Слід зазначити, що цей недолік характерний для всіх програм для копіювання веб-сайтів.

Wget – безкоштовна програма з відкритим програмним кодом, запускається та керується за допомогою командного рядка операційної системи, використовується для завантаження файлів з комп'ютерних мереж. Може завантажувати будь-які файли з мережі Інтернет (у тому числі і (X)HTML-сторінки) за протоколами HTTP** та HTTPS (безпеч-

---

* css – файли, що містять каскадні таблиці стилів для оформлення HTML-сторінок.

** HTTP – протокол передачі даних, що використовується в комп'ютерних мережах, основним призначенням якого є передача веб-сторінок.

ний HTTP), а також файли і списки директорій за протоколом FTP*. Файли можна завантажувати рекурсивно, як з одного сайту з визначеною глибиною слідування за посиланнями, так і з декількох. Недоліками цієї програми є те, що немає можливості встановлювати кількість одночасних з'єднань, а також відсутність можливості опрацювання посилань, сформованих за допомогою Java-скриптів.

Heritrix – веб-краулер з відкритим програмним кодом. Був створений за ініціативою неприбуткової організації, що розташована у Сан-Франциско (Каліфорнія, США), метою якої є збереження культурного надбання. Написаний мовою програмування Java. Головний інтерфейс реалізовано у вигляді веб-інтерфейсу, також має засіб для роботи за допомогою командного рядка. Цей веб-краулер входить до програмного засобу "Web Curator Tool", розробка якого ведеться з 2006 року за ініціативою Національної бібліотеки Нової Зеландії та Британської бібліотеки[11]. Остання стабільна версія (1,6), що вийшла 5 грудня 2012, доступна для вільного завантаження з офіційного веб-сайту її розробника[12]. Програмний засіб представляє собою повноцінне середовище керування процесами селективного копіювання веб-сайтів, управління цими інформаційними об'єктами, організацією доступу до них користувачів тощо.

Відділом інформаційних технологій ЦДЕА України було здійснено встановлення зазначеного засобу на тестовий майданчик та проведена його дослідна експлуатація. За результатами, отриманими від використання "Web Curator Tool", можна зробити висновок, що зазначений засіб загалом та "Heritrix", як веб-краулер, зокрема, не підходять для використання в інформаційній системі ЦДЕА України. Це пов'язано зі специфікою створення архівних копій веб-сайтів (далі – АКВС) та проведення їх технічної перевірки. Повноцінне копіювання веб-сайтів неможливе внаслідок низки технічних та організаційних причин: використання мов програмування, що обробляються на стороні клієнта при генерації посилань; використання технологій Web 2.0** під час організації зберігання контенту, наприклад, зберігання потокового відео на сторонніх ресурсах. Це відбувається тому, що помилки в програмному коді та відсутній контент визначається під час технічної перевірки. Для забезпечення повноти та коректності відтворення АКВС проводиться довантаження контенту та коригування програмного коду в ручному режимі. Heritrix зберігає веб-сайти у форматі WARC, що є архівом, який містить копію цільового веб-сайту. Така форма зберігання інфор-

---

* FTP – протокол передачі файлів, що дає можливість абоненту обмінюватися двійковими і текстовими файлами з будь-яким комп'ютером мережі.

** Web 2.0 – методика проектування систем, яка шляхом обліку мережевої взаємодії стає тим кращою, чим більше людей користується нею.

мації, на відміну від сукупності HTML-файлів веб-сторінок, графічних аудіо-, відеофайлів та каскадних таблиць стилів, що створюється веб-краулером HTTrack або програмним засобом Wget, ускладнює коригування цільового веб-сайту.

Параметри копіювання програм веб-краулерів та їх порівняння представлено у таблиці 1.

Таблиця 1

Порівняння *веб-краулерів*

	HTTrack	Wget	Heritrix
1	2	3	4
Відкритий програмний код	+	+	+
Налаштування рівнів завантаження	+	+	+
Довантаження з місця зупинки завантаження	+	+	+
Налаштування кількості одночасних з'єднань	+	–	+
Вибір типів файлів для завантаження	+	+	+
Можливість створення копій декількох сайтів одночасно	+	+	+
Закриття зовнішніх посилань	+	+	+
Формат кінцевого результату	Статичний веб-сайт у вигляді html-, css-, js-файлів з аудіо-, відео-, графічними файлами	Статичний веб-сайт у вигляді html-, css-, js-файлів з аудіо-, відео-, графічними файлами	Архів у форматі WARC

Після випробування розглянутих програм виявлено такі загальні недоліки:

– жодна з програм не може обробляти посилання, що формуються за допомогою мови програмування Java-скрипт на стороні користувача;

– ускладнене або взагалі неможливе завантаження контенту в автоматичному режимі, який закрито правовласниками для завантаження, наприклад, за допомогою заборони копіювання файлів скриптів та каскадних таблиць стилів зі службових директорій;

– жодна з програм не забезпечує в автоматичному режимі гарантування відповідності контенту завантаженого веб-сайту контенту цільового веб-сайту. До цього часу в Україні ще не побудовано жодної комплексної системи захисту інформації в автоматизованих інформаційних системах, які б у своїй роботі використовували програмне забезпечення, що розглядається в цій статті. Тому можна стверджувати, що в України жоден програмний засіб, призначений для копіювання веб-сайтів, не сертифіковано Державною службою спеціального зв'язку та захисту інформації України (далі – ДССЗЗІ). Разом з тим архіви повинні гарантувати відповідність інформаційного наповнення архівної копії веб-сайту й оригіналу веб-сайту. Зважаючи на те, що процес копіювання відбувається незахищеними публічними інформаційними мережами, теоретично можливі перехоплення та підміна інформації, що копіюється. Частково ця проблема вирішується під час експертизи цінності в ЦДЕА України створеної копії цільового веб-сайту методом порівняння її з оригіналом. Однак, безперечно, ця проблема потребує більш детального опрацювання на рівні покращення функціоналу програмних засобів, що використовуються для копіювання веб-сайтів з урахуванням потреб безпеки та подальшою сертифікацією цих програмних продуктів ДССЗЗІ.

Програма Heritrix не відповідає вимогам копіювання веб-сайтів ЦДЕА України, оскільки Heritrix зберігає файли скопійованих веб-сайтів у власних архівних файлах, що ускладнює їх подальшу технічну перевірку.

На сьогодні ЦДЕА України використовує комплексний підхід до створення АКВС – за замовчуванням використовується веб-краулер HTTrack, що в автоматичному режимі, згідно із заданими параметрами, створює локальний примірник цільового веб-сайту. У випадку, якщо технічна перевірка виявляє відсутні інформаційні об'єкти в створеному примірнику, відбувається ручне довантаження останніх за допомогою програмного засобу Wget. Але навіть таке поєднання не забезпечує ідеального копіювання веб-сайтів, тому ЦДЕА України постійно слідкує за тенденціями розвитку засобів для копіювання веб-сайтів, а також намагається впровадити альтернативний метод прийняття на архівне зберігання веб-сайтів, у формі файлів зрізів (резервних копій) бази даних та програмного комплексу, що забезпечує відтворення інформації із цієї бази даних у формі HTML-сторінок.

[1] Положення про Центральний державний електронний архів України [Електронний ресурс]. − Режим доступу: http://tsdea.archives.gov.ua/ua/base/759_21052012.pdf. – Назва з екрана.

[2] LIBIS [Електронний ресурс]. − Режим доступу: http://www.libis.lt/lang.do;jsessionid=884B278AD4B9979476578BB0245AC233?language=en. – Назва з екрана.

[3] Archives New Zealand [Електронний ресурс]. − Режим доступу: http://archives.govt.nz/advice/continuum-resource-kit/publications-publication-type. – Назва з екрана.

[4] Депонирование электронных документов во Франции: пятилетний опыт применения нового законодательства и уроки на будущее [Електронний ресурс]. – Режим доступу: http://www.gpntb.ru/ntb/ntb/2012/10/ntb_10_8_2012-%D0%BF%D0%B8%D1%82%D0%B5%D1%80-.pdf. – Назва з екрана.

[5] Foundatiton of National Scientific Computing [Електронний ресурс]. – Режим доступу: http://www.fccn.pt/en/. – Назва з екрана.

[6] Crawling the UK web domain [Електронний ресурс]. – Режим доступу: http://britishlibrary.typepad.co.uk/webarchive/2013/09/domaincrawl.html. – Назва з екрана.

[7] Helsinki City Library [Електронний ресурс]. – Режим доступу: http://www.hel.fi/hki/Kirjasto/en/. – Назва з екрана.

[8] Pandora Australia's Web Archive [Електронний ресурс]. – Режим доступу: http://pandora.nla.gov.au. – Назва з екрана.

[9] National library of China [Електронний ресурс]. – Режим доступу: http://www.nlc.gov.cn/newen. – Назва з екрана.

[10] National library of the Netherlands [Електронний ресурс]. – Режим доступу: http://www.kb.nl/en. – Назва з екрана.

[11] The British library [Електронний ресурс]. – Режим доступу: http://www.bl.uk/#. – Назва з екрана.

[12] Heritrix [Електронний ресурс]. – Режим доступу: http://crawler.archive.org/downloads.html. – Назва з екрана.

Рассматривается опыт некоторых стран о сохранении веб-сайтов. Анализируются программные средства для создания архивных копий веб-сайтов.

Ключевые слова: веб-сайт; архивное сохранение веб-сайтов; веб-краулер.

There is considered the experience of some countries on the preservation of web resources in the article. The author analyses the program means of the creating of archival copies of web sites.

Key words: web site; the archival preservation of the web sites; web crawler.