УДК 930.25:004.738.5

Т. Я. КУПРУНЕЦЬ*

АРХІВНЕ КОПІЮВАННЯ ВЕБ-САЙТІВ: ПРОБЛЕМИ ТА ШЛЯХИ ЇХ ВИРІШЕННЯ

Аналізується веб-сайт з точки зору архівознавства. Розглядається процес архівного копіювання веб-сайта. Визначаються деякі проблеми, що впливають на процес архівного копіювання веб-сайта та викликають невідповідності відтворення його архівної копії.

Ключові слова: веб-сайт; архівне копіювання веб-сайтів; проблеми копіювання веб-сайтів; гіперпосилання.

Виникнення мережі Інтернет у 90-х рр. XX ст., у тому вигляді, до якого ми звикли, відкрило суспільству новий простір створення, розповсюдження та обміну інформацією[1]. Вперше за історію людства інформація була позбавлена жорсткого прив'язування до певного носія. Перебуваючи у вигляді струму в дротах каналів зв'язку, електромагнітних хвиль у бездротових мережах, магнітного поля на магнітних стрічках та жорстких дисках, або у всіх цих станах одночасно, інформація зазнала якісних змін деяких її властивостей. Основних змін зазнала така властивість, як доступність інформації. Насамперед на це вплинула швидкість розповсюдження інформації у новому інформаційному середовищі. Завдяки сучасним технологіям інформація у такому середовищі розповсюджується миттєво.

Революційним кроком у формалізації подачі інформації у мережі Інтернет стало розміщення першого веб-сайта в 1990 році[2]. За своєю суттю веб-сайт – це сукупність сторінок, що містять текстову, графічну, аудіо-, відеоінформацію, яка структурована за допомогою стандартної мови гіпертекстової** розмітки веб-сторінок в Інтернеті HTML

---

* *Купрунець Тарас Ярославович* – головний спеціаліст відділу інформаційних технологій Центрального державного електронного архіву України.

** Гіпертекст – форма організації тексту, при якій його одиниці представлені не в лінійній послідовності, а як система явно вказаних можливих переходів (зв'язків) між ними. Користуючись цими зв'язками, можна читати матеріал в будь-якому порядку, утворюючи різні лінійні тексти. Найпростіший приклад гіпертексту "доінтернетовської епохи" – це якийсь словник чи енциклопедія, де кожна стаття має посилання до інших статей цього ж словника (енциклопедії). У результаті читати такий текст можна по-різному: від однієї статті до іншої, ігноруючи гіпертекстові посилання; читати статті одну за одною, користуючись посиланнями; переходити від одного посилання до іншого, обираючи матеріали за потребою. Загальновідомим і яскравим прикладом сучасного гіпертексту є сторінки веб-сайта. Відповідно гіпертекст містить у собі гіперпосилання. Гіперпосилання дозволяють переходити від одного (вихідного) тексту до безлічі інших текстів, розміщених у web-мережі.

(HyperText Markup Language). Тому такі сторінки ще називаються HTML-сторінками. Вони пов'язані між собою системою навігації, яка формується з унікальних адрес, що мають сторінки веб-сайта. Ці унікальні адреси називаються URL*, або, з позиції HTML, – гіперпосиланнями. Вони описуються у специфікації RFC 3986[3]. URL мають не тільки сторінки веб-сайта, але і окремі графічні, аудіо, відеоматеріали, що містяться на них. Беручи до уваги організаційну структуру веб-сайта та виходячи із визначення інформаційного ресурсу в законі України "Про Національну програму інформатизації" від 04.02.1998 № 74/98-ВР[4], веб-сайти можна віднести до електронних інформаційних ресурсів.

Веб-сайти розміщуються в мережі Інтернет за допомогою хостингу** – послуги розміщення файлів веб-сайта на публічному, постійно доступному сервері***. Взаємодія користувача з веб-сайтом відбувається за допомогою спеціального програмного забезпечення – браузера****. При наборі у браузері адреси сторінки веб-сайта він під'єднується до сервера, де фізично розміщуються файли веб-сайта, отримує від нього дані, згідно введеного гіперпосилання, і форматує їх у вигляді HTML-сторінки для представлення користувачеві або намагається викликати зовнішню програму, яка це зробить, у залежності від формату отриманих даних. HTML-сторінки містять звичайний текст і посилання на інші типи даних (аудіовізуальних документів), що

---

* URL (англ. Uniform Resource Locator – єдиний вказівник на ресурс) – стандартизована адреса певного ресурсу (такого як документ, чи зображення) в інтернеті (чи деінде). Запропонований Т. Бернерс-Лі. URL складається з назви протоколу доступу (HTTP, FTP, telnet, gopher та ін.) та описання шляху до ресурсу, формат якого залежить від схеми доступу до цього ресурсу: [[<протокол>://<сервер>[:<порт>][/<шлях>][/<файл>[#<розділ>]]. Квадратні дужки визначають, що елемент URL стає обов'язковим тільки за його наявністю.

** Хо́стинг (також інколи го́стинг, англ. hosting) – послуга, що надає дискову пам'ять для розміщення інформації клієнта на сервері. Зазвичай, терміном хостинг визначають послугу розміщення файлів веб-сайта на сервері, на якому встановлене програмне забезпечення (веб-сервер), необхідне для обробки запитів до цих файлів.

*** Сервер (англ. server – «служка») – окремий комп'ютер чи програма, головною ознакою яких є здатність машини чи програми тривало (постійно) працювати автономно, без втручання людини, реагуючи на зовнішні події відповідно до функціонального призначення встановленого програмного забезпечення. Сервер як комп'ютер – це комп'ютер у мережі, який надає користувачам свої обчислювальні, дискові (для зберігання інформації) ресурси та доступ до послуг, що забезпечує встановлене програмне забезпечення.

**** Браузер (англ. browser – переглядач), також броузер, оглядач, (веб-)переглядач – програмне забезпечення для комп'ютера або іншого електронного пристрою, як правило, під'єднаного до Інтернету, що дає можливість користувачеві взаємодіяти з текстом, малюнками або іншою інформацією на гіпертекстовій веб-сторінці, що зберігається у пам'яті пристрою, або на віддаленому сервері. Веб-переглядач за допомогою посилань дозволяє користувачеві швидко та просто отримувати інформацію, розміщену на багатьох веб-сторінках.

заключні у конструкції мови HTML. Коли браузер опрацьовує таку послідовність, то він відтворює тільки текст та аудіовізуальну інформацію, пропускаючи код HTML, але опрацьовуючи його. За допомогою цього коду браузер форматує текст і розташовує аудіовізуальну інформацію на HTML-сторінці.

Форма подачі інформації в мережі Інтернет у вигляді веб-сайтів швидко набула популярності і, фактично, стала стандартом розміщення інформації для загального доступу. За час розвитку веб-сайтів постійно удосконалюється їх структура. Від статичного набору HTML-сторінок (статичних веб-сайтів[5]) вони еволюціонували до складних систем (динамічних веб-сайтів[6]), у яких інформація зберігається у базах даних, а сторінки формуються спеціальними програмними комплексами. Ці комплекси написані такими мовами програмування, як PHP*, ASP.NET**, JAVA*** тощо, що формують HTML-сторінки виключно за запитом користувача, взаємодіючи з системами керуванням базами даних**** (далі – СКБД), наприклад PostgreSQL*****, MySQL******,

---

* PHP (англ. PHP: Hypertext Preprocessor – PHP: гіпертекстовий препроцесор, попередня назва: Personal Home Page Tools) – скриптова мова програмування, була створена для генерації HTML-сторінок на стороні веб-сервера. PHP є однією з найпоширеніших мов, що використовуються у сфері веб-розробок (разом із Java, ASP.NET). PHP підтримується переважною більшістю хостинг-провайдерів.

** ASP.NET – технологія створення веб-сайтів і веб-сервісів від компанії Майкрософт. Вона є складовою частиною платформи Microsoft. NET і розвитком старішої технології Microsoft ASP.

*** Java (вимовляється Джава; інколи – Ява) – об'єктно-орієнтована мова програмування, випущена компанією Sun Microsystems у 1995 році. Синтаксис мови багато в чому походить від мов програмування C та C++. У офіційній реалізації, Java програми компілюються у байткод, який при виконанні інтерпретується віртуальною машиною для конкретної платформи.

**** Система керування базами даних (СКБД) – комп'ютерна програма чи комплекс програм, що забезпечує користувачам можливість створення, збереження, оновлення, пошук інформації та контролю доступу в базах даних.

***** PostgreSQL (вимовляється «Пост-грес-Кью-ель», або «постгрес») – вільна об'єктно-реляційна СКБД. Є альтернативою як комерційним СКБД (Oracle Database, Microsoft SQL Server та інші), так і СКБД з відкритим кодом (MySQL, SQLite). Порівняно до інших проектів з відкритим кодом, такими як Apache, FreeBSD або MySQL, PostgreSQL не контролюється якоюсь однією компанією, її розробка можлива завдяки співпраці багатьох людей та компаній, які хочуть використовувати цю СКБД та впроваджувати у неї найновіші досягнення.

****** MySQL – вільна СКБД. Розробку й підтримку MySQL здійснює корпорація Oracle, що отримала права на торгову марку разом з поглинутою Sun Microsystems, яка раніше придбала шведську компанію MySQL AB. Продукт розповсюджується як за відкритою ліцензійною угодою GNU (GNU

Microsoft SQL Server*, Oracle Database**. За визначенням ст. 1 Закону України "Про Національну програму інформатизації" база даних – це "іменована сукупність даних, що відображає стан об'єктів та їх відношень у визначеній предметній галузі", а база знань – це "масив інформації у формі, придатній до логічної і смислової обробки відповідними програмними засобами"[7].

Популярність електронного середовища як інформаційного майданчика призвела до того, що його користувачі почали створювати документи виключно в електронній формі. Кількість нових веб-сайтів постійно зростає, однак разом з цим відбувається зникнення вже існуючих[8]. Ігнорувати цю цифрову спадщину людства, зокрема України, не можна[9], інакше вона буде втрачена. Тому в розвинених країнах були створені профільні організації, що відповідають за збереження Інтернет-ресурсів, що мають історичну, культурну та наукову цінність. У кожній країні обираються власні принципи відбору веб-сайтів для їх передачі на постійне зберігання. Ці принципи формують певні загальні підходи до збереження веб-сайтів. Серед них можна виділити два основних:

– копіювання всіх веб-сайтів у окремій доменній зоні***. Наприклад, таким шляхом пішла Нова Зеландія, де у 2003 р. на законодавчому рівні[10] Національна бібліотека Нової Зеландії отримала право на копіювання та зберігання всіх електронних публікацій та Інтернет сайтів країни. Було створено Національний архів цифрової спадщини (National Digital Heritage Archive, NDHA);

– відбір тематичних колекцій. Цим шляхом йде Україна, зокрема Центральний державний електронний архів України, де створюються

General Public License), так і за власною комерційною ліцензією. Окрім цього розробники створюють функціональність на замовлення ліцензованих користувачів.

* Microsoft SQL Server – комерційна СКБД, що розповсюджується корпорацією Microsoft. Мова Transact-SQL, що використовується для запитів до даних у базах, створена спільно компаніями Microsoft та Sybase. Transact-SQL є реалізацією стандарту ANSI/ISO щодо структурованої мови запитів SQL (Structured Query Language) із розширеннями. Використовується як для невеликих і середніх за розміром баз даних, так і для великих баз даних масштабу підприємства. Багато років вдало конкурує з іншими СКБД.

** Oracle Database (часто просто Oracle) – об'єктно-реляційна СКБД від компанії Oracle Corporation. Випускається з 1977 року та є однією з лідерів комерційних СКБД.

*** Доменна зона – сукупність доменних імен певного рівня, що входять в конкретний домен. Наприклад, доменна зона com.ua включає всі доменні імена третього рівня в цьому піддомені. Наприклад: домен d.com.ua, домен domain.com.ua або domainname.com.ua – це все домени піддомену .com.ua, або домени, що належать доменній зоні .com.ua

колекції веб-сайтів, присвячені чорнобильській катастрофі, виборам у країні, проведенню чемпіонату Європи з футболу "Євро-2012" тощо.

Обрання одного з цих підходів для поповнення архівних фондів веб-сайтами напряму залежить від правового поля кожної конкретної держави. Адже простого бажання профільних організацій створювати копії окремих доменних зон або окремих веб-сайтів та об'єднувати їх у тематичні колекції не достатньо. Вони повинні мати право, закріплене на законодавчому рівні, проводити таку діяльність[11].

У випадку країн, де процес законодавчого закріплення права на копіювання окремих доменних зон профільними організаціями ще не завершився (наприклад, у Нідерландах, Чехії та Україні)[12], можливе використання лише другого методу. Відбір веб-сайтів для включення до тематичних колекцій, їх копіювання, подальшого зберігання та надання доступу до збережених копій користувачам проводиться на основі договорів, укладених з правовласниками веб-сайтів, зокрема в Україні у межах ініціативного документування. Це призводить до того, що існує ймовірність втрати унікальних веб-сайтів, з позиції їх історичної, культурної або наукової цінності. Причиною цього може стати відмова правовласника надати свій веб-сайт для передачі на постійне зберігання. Це проблема виключно правового характеру.

З позиції документознавства веб-сайт – це нове явище. Веб-сайт є постійно змінюваним інформаційним об'єктом. У звичному для архівістів сенсі, за аналогією зі звичайними документами, оригіналом веб-сайта вважається інформаційний об'єкт, що було отримано після завершення супроводу веб-сайта. Зважаючи на те, що зміни веб-сайтів відбуваються таким чином, що значна частина інформації за результатами цих змін може бути повністю оновлена, тобто втрачена, у світі існує практика копіювання веб-сайтів станом на певний проміжок часу. З певною ймовірністю такий підхід забезпечує збереження усього обсягу інформації, що містив веб-сайт. Головним чинником тут є визначення оптимальної періодичності копіювання веб-сайта. Результатом подібного періодичного копіювання є масив архівних копій веб-сайта, кожна з яких містить інформацію станом на момент її створення.

Проблеми архівного копіювання веб-сайтів не закінчуються сферою правового регулювання вищезазначеного процесу. На етапі копіювання можуть виникати обставини, які здатні призвести до невідповідності відтвореної архівної копії веб-сайта з веб-сайтом на сервері правовласника в Інтернеті. Повна відсутність деяких HTML-сторінок в архівній копії веб-сайта, або ж часткова відсутність елементів HTML-сторінок можуть бути наслідком наступних дій:

- розривами з'єднання з Інтернетом;
- недоступністю серверу, що надає хостинг для веб-сайта, що копіюється (далі – цільового веб-сайта);

- неправильне розгортання та налаштування програмного забезпечення копіювання веб-сайтів;
- особливості програмного коду сторінок веб-сайта, які відповідають за механізм формування гіперпосилань.

Якщо проблеми Інтернетз'єднання та доступності сервера – це зона відповідальності третіх сторін (Інтернет- та хостинг-провайдерів*), то програмне забезпечення для копіювання веб-сайтів та якість програмного коду веб-сайтів – це вже зона відповідальності профільної організації, що здійснює копіювання, та правовласника цільового веб-сайта відповідно. Останній фактор потребує окремої уваги, адже він не залежить від сторони, що здійснює копіювання веб-сайта, а тому його подолання може або викликати найбільше складнощів, або взагалі унеможливити повноцінне копіювання веб-сайта.

Для розуміння причин виникнення зазначеної проблеми потрібно детальніше розглянути сам процес архівного копіювання веб-сайтів. Зазвичай виокремлюють два способи здійснення цього процесу.

Перший спосіб полягає у копіюванні з серверу, що надає послуги хостингу, при підключенні безпосередньо до його файлової системи**. У випадку статичного веб-сайта, це копіювання його HTML-сторінок та супровідних файлів. Супровідними файлами веб-сайта виступають файли каскадних таблиць стилів***, що відповідають за візуальне оформлення, файли скриптів (наприклад мовою JavaScript****), що відповідають за розширення функціоналу веб-сайта та текстові й ауді-овізуальні матеріали, що використовувались при створенні та необхідні

---

* Провайдер послуг Інтернету, також Інтернет-провайдер, (від Internet Service Provider – ISP; англ. to provide – забезпечувати, надавати доступ) – організація, яка надає послуги доступу та передачі (інформації) певними інформаційними каналами.

** Файлова система — спосіб організації даних, який використовується операційною системою для збереження інформації у формі файлів на носіях інформації. Також цим поняттям позначають сукупність файлів та директорій (каталогів, тек), які розміщуються на логічному або фізичному диску.

*** Каскадні таблиці стилів (англ. Cascading Style Sheets – CSS) – спеціальна мова, що використовується для відображення сторінок, написаних мовами розмітки даних. Найчастіше CSS використовують для відтворення інформації, що містять сторінки написані з використанням мов HTML та XHTML.

**** JavaScript – назва реалізації стандарту мови програмування ECMAScript компанії Netscape. Найпоширеніше і найвідоміше застосування мови – написання сценаріїв для веб-сторінок. На сьогоднішній день підтримується більшістю браузерів. Текст програми включається безпосередньо в HTML-документ і інтерпретується браузером (точніше, вбудованим у браузер рушієм JavaScript). Найчастіше застосовується для часткової автоматизації обробки і маніпуляції даними HTML-сторінки.

для повноцінного його функціонування. У випадку динамічного веб-сайта – це копіювання програмного комплексу і бази даних, що формують HTML-сторінки веб-сайта, а також супровідних файлів.

Вищезазначений метод, з огляду на те, що копіювання відбувається безпосередньо з файлової системи сервера, забезпечує 100% гарантію копіювання веб-сайта без помилок. Однак при розгортанні такої копії на власному сервері для надання до нього доступу користувачам необхідно врахувати, що сервер має відповідати умовам, за яких функціонував веб-сайт на сервері правовласника в Інтернеті. Для статичних веб-сайтів виконання цієї умови не викликає особливих складнощів, адже достатньо встановити веб-сервер* щоб архівна копія веб-сайта стала доступна в мережі. Проте у випадку динамічних веб-сайтів потрібно відповідне середовище виконання програм, яке б забезпечило правильне функціонування скопійованого програмного комплексу та СКБД для забезпечення доступу до копії бази даних, з якої будуть формуватись HTML-сторінки. Власниками веб-сайтів часто використовуються комерційні середовища та СКБД, що ускладнюють накопичення та розгортання копій веб-сайтів у такий спосіб.

Другий спосіб копіювання веб-сайтів – це копіювання безпосередньо з мережі Інтернет за допомогою спеціальних програмних засобів. Цей процес ще називають веб-харвестингом**, а програмні засоби, що використовуються для отримання копій веб-сайтів з мережі Інтернет, – веб-краулерами***. Краулер, імітуючи браузер, звертається за визначеним гіперпосиланням до відповідної HTML-сторінки веб-сайта, копіює її, сканує наявність гіперпосилань на складові її змісту, супровідні файли та наступні HTML-сторінки і переходить за цими гіперпосиланнями, повторюючи свої дії уже стосовно наступних складових веб-сайта. Так продовжується доти, доки не буде скопійовано HTML-сторінку за останнім гіперпосиланням[13].

Слід зазначити різницю в отриманих результатах при копіюванні статичних та динамічних веб-сайтів цим способом. При копіюванні

---

* Веб-сервер (англ. Web Server) – це програмне забезпечення, що встановлюється на сервер та приймає запити від клієнтів, зазвичай браузерів, надсилає їм відповіді, зазвичай у формі HTML-сторінок, зображень, файлів різних форматів, медіа-потоків або інших даних.

** Веб-харвестинг (від англ. harvest «збирати врожай») – процес копіювання веб-сайта, або сайтів з мережі Інтернет, що здійснюється за допомогою спеціалізованого програмного забезпечення. Може бути ініційований як для одного веб-сайта, так і для множини, наприклад, цілої доменної зони.

*** Веб-краулер (англ. Web-crawler, «веб-паук», краулер) – програма, що призначена для перебору HTML-сторінок веб-сайта з метою їх подальшого копіювання. За принципом дії краулер імітує звичайний браузер. Він аналізує зміст HTML-сторінки, зберігає його і переходить по посиланням на наступні сторінки.

веб-сайтів, які мають статичну структуру, архівні копії будуть ідентичні джерелам. Однак у випадку з динамічними веб-сайтами архівні копії та джерела будуть відрізнятися. Це обумовлено тим, що краулер, імітуючи браузер, отримує за кожним гіперпосилання, за яким звертається до веб-сайта, або вже готову HTML-сторінку, згенеровану програмним комплексом власника веб-сайта, або супровідні файли. Краулер не має прямого доступу до програмного забезпечення та бази даних динамічного веб-сайта, тому результатом його роботи буде статична копія динамічного веб-сайта.

Перевагами другого способу копіювання веб-сайтів є:

- відсутність необхідності прямого підключення до файлової системи серверу, що надає хостинг для цільового веб-сайта;
- результатом копіювання є статичний веб-сайт, для забезпечення доступу користувачів до якого не потрібно встановлювати та налаштовувати на власному сервері додаткове комерційне програмне забезпечення.

Виходячи із вищезазначених переваг, цей спосіб є основним, для копіювання веб-сайтів. Однак саме з ним пов'язані проблеми при копіюванні, що викликані особливостями програмного коду HTML-сторінок веб-сайта, що відповідають за формування гіперпосилань.

Як зазначалось вище, ключовим параметром роботи краулера є гіперпосилання. Переходячи по них він копіює відповідні HTML-сторінки та супровідні файли. Якщо гіперпосилання на веб-сайті будуть у формі, що опрацьовується краулером некоректно або не опрацьовується взагалі, інформація, що асоціюється з цими гіперпосиланнями, скопійована не буде. Саме тому механізм формування гіперпосилань є дуже важливим елементом у процесі копіювання веб-сайтів.

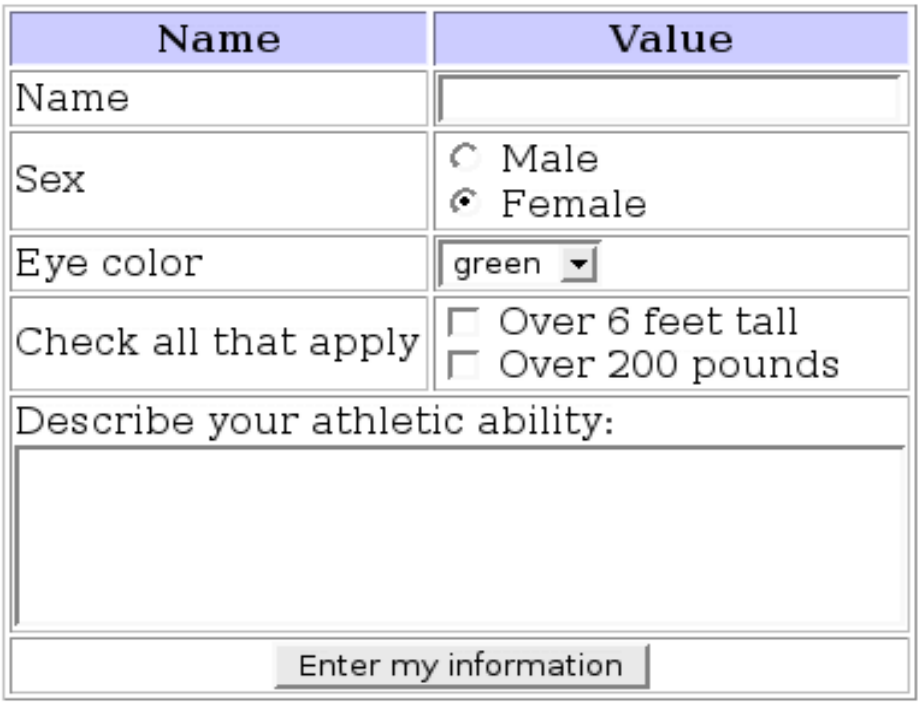
Гіперпосилання можуть формуватись декількома способами[14]:

- визначатися остаточно в HTML-сторінках на етапі створення веб-сайта – статичні гіперпосилання;
- генеруватися на стороні серверу разом з HTML-сторінкою у випадку динамічного веб-сайта;
- створюватися за допомогою веб-форм*;
- генеруватися на стороні користувача, наприклад, мовою програмування JavaScript.

Гіперпосилання, статично прописані в HTML-сторінці або згенеровані в ній з боку сервера, не викликають труднощів при роботі з ними краулера. Тоді як формування гіперпосилань за допомогою веб-форм та мови JavaScript взагалі виключає можливість повноцінного копіювання веб-сайта. Розглянемо детальніше зазначену проблему.

---

* Веб-форма (форма) – елемент веб-сторінки, що дає користувачам можливість вводити інформацію і відправляти її на сервер для подальшої обробки. В окремих випадках може використовуватись для побудови посилань.



Стосовно використання веб-форм при формуванні гіперпосилань на веб-сайтах та їх впливу на продуктивність роботи краулера потрібно звернути увагу на наступне. Зазвичай веб-форми використовують для інтерактивної взаємодії користувача з веб-сайтом[15]. Вони можуть мати текстові поля, які заповнюються користувачем, кнопки вибору та випадаючі списки, які дозволяють користувачу обрати один або декілька із запропонованих варіантів (див. малюнок).

Після натискання на кнопку підтвердження дані, що були внесені користувачем, передаються на сервер. Передача здійснюється за гіперпосиланням, вказаному в спеціальному параметрі веб-форми. Залежно від вказаних даних користувач отримує у свій браузер відповідь у формі нової HTML-сторінки. Дані, які були введені в текстові поля або обрані у вигляді варіантів, попередньо визначених розробником вебсайта, наприклад, у вигляді випадаючого списку, виступають у ролі параметрів. На їх основі формується HTML-сторінка з відповіддю. Існує два способи передачі цих параметрів веб-формами – GET та POST. Їх відмінність полягає у тому, що метод GET додає введені дані безпосередньо в гіперпосилання, за яким іде звернення до програми на сервері, що опрацьовує таке гіперпосилання, а метод POST – у тіло звернення[16]. Параметри, що передаються методом GET, видно в адресному рядку браузера, параметри, що передаються методом POST, в адресному рядку браузера не видно. Завдяки цьому принципу, звертання до сторінок веб-сайта за допомогою веб-форм з передачею параметрів методом

POST використовується тоді, коли не бажано, щоб параметри були доступні користувачу, наприклад в цілях безпеки. У той же час, краулер, копіюючи веб-сайт, не буде емулювати натискання підтверджуючої кнопки. Як наслідок цього, HTML-сторінка з відповіддю не буде згенерована та скопійована. Це ж стосується веб-форм, що передають параметри за методом GET.

Аналізуючи вплив гіперпосилань, що генеруються мовою програмування JavaScript, на результат роботи краулера потрібно відзначити той факт, що використання при побудові веб-сайтів вищезазначеної технології вже набуло значної популярності і вона продовжує зростати[17]. Поясненням цьому є принцип її роботи. Програмний код, створений нею, на відміну від серверних мов програмування, виконується з боку користувача, а не сервера. Наслідком цього є розвантаження сервера, що обслуговує веб-сайт. Програмний код JavaScript виконується без перезавантаження HTML-сторінки, що забезпечує покращення показників швидкості роботи та динаміки веб-сайта. Однак переваги, які забезпечили популярність цієї мови програмування, є причинами, що перешкоджають автоматичному створенню повної архівної копії веб-сайта. Це пояснюється тим, що гіперпосилання, які формуються мовою JavaScript перебувають в "розібраному" стані, тобто сервер, передаючи HTML-сторінку користувачу, не формує з коду JavaScript готові гіперпосилання, а покладає цю роботу на браузер. Найчастіше поштовхом для формування браузером гіперпосилань є якась ситуація[18], це може бути натискання кнопки, у вигляді якої реалізоване гіперпосилання, клік по зображенню тощо. Найчастіше такі прийоми реалізації гіперпосилань використовуються при створенні допоміжних навігаційних панелей, веб-галерей з різними зображеннями, слайдерів, що автоматично прокручують інформацію у певній області HTML-сторінки. У звичайному режимі роботи користувача з веб-сайтом такий принцип не викликає труднощів, адже веб-сайт в повному обсязі знаходиться на сервері та готовий видати інформацію з будь-якого, згенерованого JavaScript, гіперпосилання. Проте під час копіювання веб-сайта краулер не може імітувати дії користувача, що призводять до створення гіперпосилань. Внаслідок цього частина гіперпосилань, за генерацію яких відповідає JavaScript, не створюються. Якщо гіперпосилання не згенеровано, не відбудеться і копіювання інформації, що доступна за цим гіперпосиланням, а отже, веб-сайт буде скопійовано не повністю. Разом з процесом підвищення популярності мови програмування JavaScript проблеми, що вона створює при копіюванні веб-сайта, теж набувають все більшого поширення.

Розглянуті вище проблеми особливо гостро постають, коли потрібно працювати не з окремо обраними веб-сайтами під час формування нечисленних архівних колекцій, а при здійсненні автоматичного масового копіювання. Наприклад, створення копій веб-сайтів національного

домену країни, часових зрізів веб-сайтів державних установ тощо. Усунення проблем, викликаних особливостями програмного коду, потребує індивідуального аналізу та ручного втручання, що унеможливлює процес автоматичного копіювання веб-сайтів.

Вирішення проблем як правового, так і технічного характеру, описаних у цій статті, загалом, не досягається тільки силами профільних організацій, що здійснюють копіювання веб-сайтів.

Одним із варіантів вирішення цієї проблеми може бути включення веб-сайтів національного домену в закон про обов'язковий екземпляр таких інформаційних об'єктів, що можуть бути відібрані для постійного зберігання, залежно від свого статусу, передаватися або до бібліотек, або до архівів, за прикладом Данії, Литви, Нової Зеландії, Норвегії тощо. Надання профільним організаціям права проводити копіювання ресурсів національного домену дозволить уникнути небезпеки втрати для дослідників та науковців майбутнього веб-сайтів, як джерел, що містять цінну історичну, культурну та науково-технічну інформацію.

У випадку з проблемами технічного характеру, їх вирішення не можливе без взаємодії з правовласниками веб-сайтів. На етапі створення веб-сайта необхідно знайти відповідь на питання, чи буде він піддаватися архівному копіюванню. У випадку позитивної відповіді необхідно вже на етапі його розробки врахувати особливості взаємодії того чи іншого програмного рішення, що використовується для створення та подальшого функціонування веб-сайта. Для цього слід ініціювати на рівні Державної архівної служби України визначення переліку типових веб-сайтів, що створюються у процесі діяльності органів державної влади та місцевого самоврядування, підприємства, установи та організації будь-якої форми власності (далі – фондоутворювачі), із зазначенням строків їх зберігання, а також вимог до структури і змісту архівних копій веб-сайтів. Зазначені документи мають бути затверджені як нормативно-правові акти, дія яких поширюється на всіх фондоутворювачів.

Разом з тим потрібно продовжувати дослідження програмного забезпечення, що використовується для копіювання веб-сайтів, та у напрямі вдосконалення цього процесу в цілому.

---

[1] *Кудрявцева С. П., Колос В. В.* Міжнародна інформація. Навчальний посібник для студентів вищих навчальних закладів. − К.: Видавничий Дім "Слово", 2005. − 400 с.

[2] The website of the world's first-ever web server [Електронний ресурс]. − Режим доступу: http://info.cern.ch/. − Назва з екрана.

[3] Uniform Resource Identifier (URI): Generic Syntax [Електронний ресурс]. − Режим доступу: http://www.ietf.org/rfc/rfc3986.txt − Назва з екрана.

[4] Про Національну програму інформатизації : Закон України від 04.02.1998 № 74/98-ВР // Відомості Верховної Ради України. – К., 1998 – № 27. – Ст. 181.

[5] Static web page [Електронний ресурс]. − Режим доступу: http://en.wikipedia.org/wiki/Static_web_page − Назва з екрана.

[6] Dynamic web page [Електронний ресурс]. − Режим доступу: http://en.wikipedia.org/wiki/Dynamic_web_page − Назва з екрана.

[7] Про Національну програму інформатизації : Закон України від 04.02.1998 № 74/98-ВР // Відомості Верховної Ради України. – К., 1998 – № 27. – Ст. 181.

[8] Peter Lyman. Archiving the World Wide Web // Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving. − Washington, Council on Library and Information, Library of Congress, April 2002. − С. 38−52.

[9] Хартия о сохранении цифрового наследия [Електронний ресурс]. − Режим доступу: http://www.un.org/ru/documents/decl_conv/conventions/digital_heritage_charter.shtml − Назва з екрана.

[10] National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003 [Електронний ресурс]. − Режим доступу: http://www.legislation.govt.nz/act/public/2003/0019/latest/DLM191962.html − Назва з екрана.

[11] *Браккер Н. В., Куйбышев Л. А.* Сбор и архивирование сетевых ресурсов. Опыт национальных библиотек зарубежных стран [Електронний ресурс]. − Режим доступу: http://www.minervaplus.ru/publish/Harvesting_Preservation_Net_Resources.doc − Назва з екрана.

[12] Там само.

[13] Web crawler [Електронний ресурс]. − Режим доступу: http://en.wikipedia.org/wiki/Web_crawler − Назва з екрана.

[14] Организация ссылок. Всевозможные оглавления. [Електронний ресурс]. − Режим доступу: http://webdesign.site3k.net/?/conjuncture/append/d/menus.html − Назва з екрана.

[15] *Молли Э.* Хольцшлаг. Использование HTML 4, 6 изд. − Вильямс, 2000. − 1008 с.

[16] GET и POST HTTP-запросы. Передача параметров в HTTP-запросах [Електронний ресурс]. − Режим доступу: http://www.myfirstsite.ru/articles/get-and-post − Назва з екрана.

[17] Douglas Crockford The World’s Most Misunderstood Programming Language Has Become the World’s Most Popular Programming Language [Електронний ресурс] // 2008. − Режим доступу: http://javascript.crockford.com/popular.html − Назва з екрана.

[18] События [Електронний ресурс]. − Режим доступу: http://javascript.ru/tutorial/events − Назва з екрана.

Анализируется веб-сайт с точки зрения архивоведения. Рассматривается процесс архивного копирования веб-сайта. Определяются некоторые проблемы, которые влияют на процесс архивного копирования веб-сайта и вызывают несоответствия воссоздания его архивной копии.

Ключевые слова: веб-сайт; архивное копирование веб-сайтов; проблемы копирования веб-сайтов; гиперссылка.

There is analyzed the web-site on the point of view of archival science in the article. The author considered the process of archival copying of web-site and indicates some problems that influence on the process of archival copying and cause the inconsistencies in the reproduction of the archival copy of the web-site.

Keywords: the web-site; the archival copying of web-sites; the problems of the copying process of web-sites; hyperlinks.