УДК 004.67

БЕРЕЗІН Б.О., науковий співробітник,
Інститут проблем реєстрації інформації НАН України

МОДЕЛЮВАННЯ ДОВГОТЕРМІНОВОГО ЗБЕРІГАННЯ ПРАВОВОЇ ІНФОРМАЦІЇ

Анотація. *Розглядаються особливості довготермінового зберігання правової інформації. Розроблено моделі загроз та негативних впливів при довготерміновому зберіганні, методи планування зберігання, які забезпечують живучість інформаційних об’єктів.*

Ключові слова: *правова інформація, довготермінове зберігання, моделі загроз, планування зберігання, живучість інформаційних об’єктів, степеневий розподіл.*

Аннотация. *Рассматриваются особенности долгосрочного хранения правовой информации. Разработаны модели угроз и негативных влияний при долгосрочном хранении, методы планирования хранения, которые обеспечивают живучесть информационных объектов.*

Ключевые слова: *правовая информация, долговременное хранение, модели угроз, планирование хранения, живучесть информационных объектов, степенное распределение.*

Summary. *The features of digital preservation of legal information are analyzed. The models of threats and negative influencing at long-term storage are developed, the methods of planning of storage, which provide vitality of information holding object.*

Keywords*: legal information, digital preservation, the models of threats, planning of storage, survivability of information objects, power-series distribution.*

Постановка проблеми та аналіз публікацій. За даними IDC, обсяги інформації, що створювались у світі за останні роки, становили у 2010 р. – 1,15 зеттабайт, а у 2012 р. вже 2,8 зеттабайт. Зростання загального обсягу інформації, що зберігається, веде до зростання обсягів даних, які повинні зберігатися довготерміново. Одна з перших у світі довідково-правових систем Lexis почала розроблятися наприкінці 60-х років минулого сторіччя в США. Зараз це одна з найбільших у світі баз даних правової інформації LexisNexis, яка надає доступ до мільярдів документів з більш як 45 тис. правових, новинних та бізнес-джерел. LexisNexis охоплює публікації з правової інформації, починаючи, з XIX сторіччя, у США, Великобританії, Канаді та інших країнах [1]. До числа найбільших баз даних правової інформації також відносять Westlaw, яка об’єднує більше ніж 40 тис. баз даних, та HeinOnline, особливістю якої є представлення документів тільки у вигляді PDF-файлів, відсканованих з першоджерел.

Все частіше в США уряди штатів публікують закони, положення про органи та установи, нормативно-правові акти органів виконавчої влади та судові накази і рішення в Інтернеті. У деяких штатах важливі правові матеріали рівня штату більше не публікуються в друкованому вигляді і доступні тільки в глобальній мережі. Для регламентації процедур забезпечення автентичності, довготермінового зберігання та доступності матеріалів через п’ятдесят, сто років штатами приймається “Типовий закон про правові акти, що публікуються в електронному вигляді” (“Uniform Electronic Legal Material Act” – UELMA) [2]. Якщо штат зберігає правові матеріали в електронному вигляді, він повинен забезпечити їх резервне копіювання і відновлення, а також цілісність матеріалів та їх постійну придатність до використання. UELMA не вимагає застосування будь-яких технологій, залишаючи вибір технологій для аутентифікації і

забезпечення збереження на розсуд штатів. Гнучкість закону, який дозволяє штатам вибрати технологію, що забезпечує отримання встановлених кінцевих результатів, дає кожному штату можливість підібрати для себе найкращий і найбільш економічно ефективний метод. Крім того, такий гнучкий і орієнтований на кінцевий результат підхід враховує те, що технології будуть з часом змінюватися; ні в який момент часу закон не “прив’язує” штат до якоїсь конкретної технології.

Один з напрямів довготермінового зберігання правової інформації пов’язаний з забезпеченням постійного доступу до інформації, що створюється тільки в цифровому вигляді (без друкованої копії) [3 – 5]. В роботі [3] наголошується на високому ризику втрати правової інформації, що створюється у вигляді веб-сторінок (публікації в електронних журналах, на блогах тощо). Це пов’язано з тим, що більшість проектів зберігання інформації спрямовано на оцифрування друкованих документів. Представлено проект Chesapeake Prtoject, запроваджений кількома правовими бібліотеками США для збору та зберігання правової інформації, доступної на веб-ресурсах, з метою включення її до національних програм зберігання.

Робота [4] присвячена дослідженням стабільності URL – тобто доступності посилань на веб-ресурси з правової інформації з плином часу. В одному з цих досліджень для набору з близько 600 веб-ресурсів (відібраних з метою довготермінового зберігання в рамках Chesapeake Prtoject) аналізувалась доступність відповідних URL-посилань в Інтернеті. В результаті виявилось, що протягом першого року стали недоступними більш як 8 % URL, за другий рік кількість недоступних URL зросла до більш як 14 %, на третьому році недоступних посилань стало понад 28 %. Таке зростання кількості недоступних URL підтверджує ризики втрати правової інформації, що створюється у вигляді веб-сторінок.

В роботі [5] розглядається заява ряду правових бібліотек про відкритий доступ до цифрових матеріалів правової освіти та припинення публікації правових видань у друкованому вигляді. Для забезпечення довготермінового доступу запропоновано оцінити наступні альтернативні рішення: архів правової інформації, заснований у 2010 р.; зберігання правового контенту в базах HeinOnline, LexisNexis та Westlaw; використання програмного забезпечення електронних архівів, таких як Portico та LOCCKS; використання можливостей Бібліотеки Конгресу США, яка приймає копії усіх правових журналів у друкованому або електронному форматі; створення інституційних репозитаріїв.

Таким чином, проведений аналіз показує актуальність рішень для забезпечення довготермінового зберігання правової інформації та її доступності, а також зменшення витрат. Світові тенденції останніх років полягають у тому, що для вирішення цієї проблеми і зменшення витрат на зберігання недостатньо розвитку традиційного апаратного і програмного забезпечення. Необхідне створення нових засобів – математичних моделей зберігання і побудованих на їх основі інструментальних засобів, програмних пакетів для вибору стратегій, правил для планування та оптимізації процесу зберігання.

Ця тенденція проявилася ще в моделі відкритої архівної інформаційної системи, рекомендованої міжнародним стандартом (Reference Model for an Open Archival Information System – OAIS). Відповідно до моделі при довготерміновому зберіганні даних необхідно враховувати вплив зміни технологій, підтримку нових видів носіїв та форматів, зміну спільнот користувачів тощо. Стандарт забезпечує основу для порівняння різних стратегій та технологій довготермінового зберігання. Серед функцій OAIS передбачається функція планування зберігання, яка забезпечує моніторинг середовища архівного зберігання, рекомендації та плани зберігання для гарантії того, щоб інформація, яка зберігається в OAIS, залишалася доступною та зрозумілою для

користувачів у довготерміновій перспективі, навіть якщо обчислювальне середовище застаріє. Функції планування зберігання включають оцінку контенту архіву та періодичні рекомендації щодо оновлення архівної інформації, рекомендації по міграції поточних запасів архіву, розробку рекомендацій стосовно архівних стандартів та політиків, забезпечення періодичних звітів з аналізу ризиків та моніторингу змін в технологічному середовищі та вимог до обслуговування користувачів.

Серед моделей, які створюються для обґрунтування планування процесів зберігання даних, можна виділити аналітичні та імітаційні моделі. Зокрема, модель, запропонована в [6], присвячена особливості довготермінового зберігання великих обсягів даних, пов’язана з необхідністю міграції даних, обміну даними через старіння форматів тощо. При плануванні зберігання великих обсягів даних необхідно враховувати час, що витрачається на міграцію даних, оскільки він може бути значним. Інструментальний засіб ReproSim призначений для імітаційного моделювання оцінки цифрових репозиторіїв з плином часу, розроблено у Віденському університеті технологій [7].

Метою статті є дослідження загроз, негативних впливів на живучість інформаційних об’єктів при довготерміновому зберіганні для розробки методів планування довготермінового зберігання.

Виклад основних положень. Особливість запропонованого підходу [8, 9] полягає в тому, що при плануванні зберігання з метою забезпечення доступності розглядається живучість інформаційних об’єктів (далі – ІО), тобто властивість виконувати основні функції в умовах негативних впливів (далі – НВ), тимчасово відмовляючись від виконання деяких другорядних функцій [10].

До основних НВ (загроз) при довготерміновому зберіганні відносять: відмови обладнання; старіння програмного забезпечення (ПЗ), форматів, обладнання; атаки; помилки операторів; катастрофи; економічні помилки і т. ін. Для підвищення живучості ІО в даній роботі досліджуються закономірності, будуються моделі різних видів НВ, загроз: множинних відмов, стану обчислювальних ресурсів, помилок на носіях даних, старіння ПО/форматів, мережевих атак [8, 9, 11] .

Моделювання множинних відмов. Для аналізу впливу близьких у часі відмов на живучість ІО у розподілених мережах зберігання даних було розроблено імітаційну модель множинних відмов [9]. Близькі за часом відмови у великій кількості вузлів можуть зменшити ефективність реплікації і, відповідно, живучість ІО. Характеристики корельованих відмов аналізувалися за допомогою вікна спостереження (часового вікна). Результати показують, що при експоненційному розподілі відмов більшість часових вікон припадає на вікна з максимальним значенням часу спостереження (Рис. 1.), а при ступеневому розподілі – на вікна з малим значенням часу (Рис. 2). Вікна з малим значенням часу спостереження (в які потрапляють близькі у часі відмови) і відповідні їм значення кількості близьких у часі відмов (а також відповідні кількості вікон) характеризують найбільш складні для забезпечення доступності даних та живучості ІО періоди.

Моделювання стану обчислювальних ресурсів у розподілених комп’ютерних системах. Для надійного функціонування у складі розподілених комп’ютерних систем передбачаються засоби моніторингу стану обчислювальних ресурсів. Ця інформація може відображати нормальний стан ресурсів; стан, що потребує уваги; критичний стан. Крім того, може бути представлена більш детальна інформація про результати виконання окремих тестів у процесі моніторингу. Модель стану обчислювальних ресурсів у розподілених комп’ютерних системах може використовуватися для опосередкованої оцінки загроз, негативних впливів на інформаційні об’єкти, що зберігаються в таких системах, планування довготермінового зберігання та забезпечення живучості. З цією

метою розробляються програмні засоби накопичення результатів моніторингу для їх подальшого аналізу.

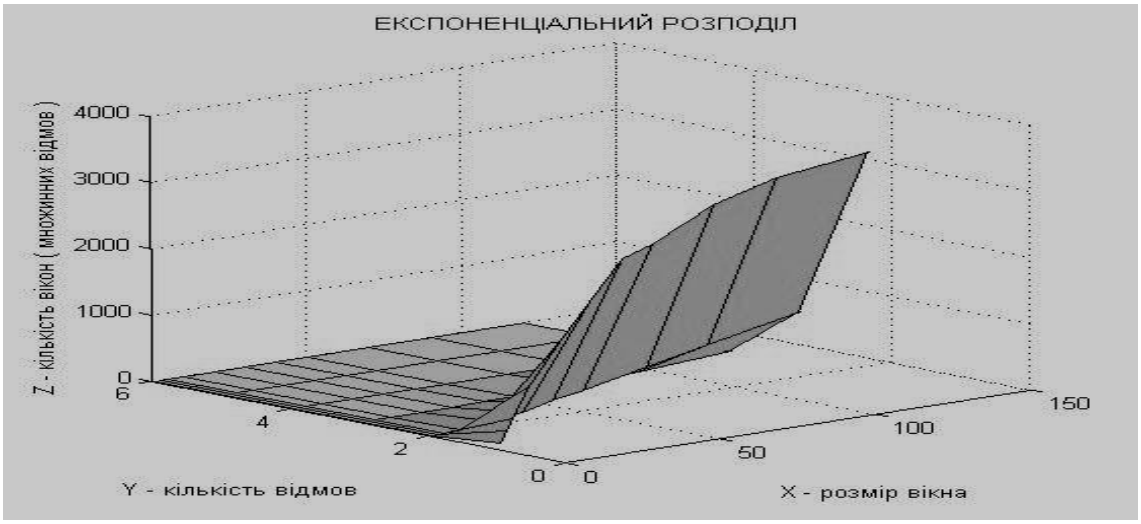

Рис. 1. Оцінка кількості множинних відмов при експоненційному розподілі.

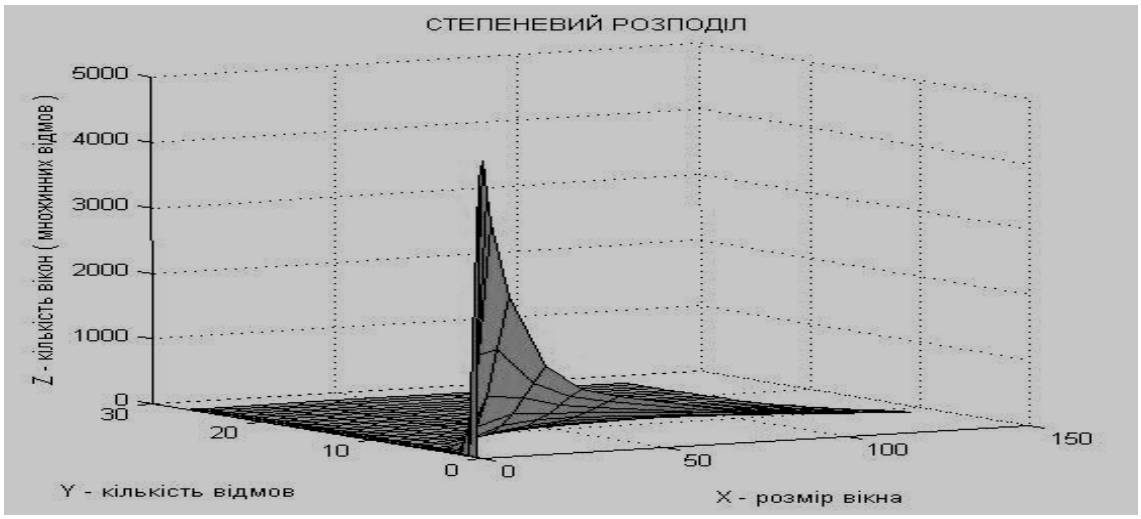

Рис. 2. Оцінка кількості множинних відмов при ступеневому розподілі.

Моделювання відмов на носіях даних. Для дослідження живучості ІО при довготерміновому зберіганні на носіях даних було зібрано статистику на основі показника помилок PI Sum 8 для DVD-дисків [11]. Дані були проранжирувані за кількістю помилок та апроксимовані з допомогою ступеневої функції. Цей та інші отримані результати обґрунтовують можливість використання моделі із ступеневим розподілом помилок.

Моделювання мережевих атак. При розробці моделі в якості опосередкованої оцінки статистики мережевих атак при довготерміновому зберіганні даних у розподілених мережах використовувалася статистика повідомлень про кібератаки, зібрана в новинах Інтернет-ресурсів. Для оцінки загроз, що створюються мережевими атаками, в якості емпіричних даних моделі використовувались результати пошуку по ретроспективній базі Рунета, створеної за допомогою технології моніторингу новин системи InfoStream. За період 2010 – 2013 рр. отримано понад півтори тисячі значень повідомлень про кібератаки. Розглядається апроксимація розподілу дат, ранжируваних за кількістю повідомлень про кібератаки за допомогою степеневої функції.

Моделювання старіння ПЗ/форматів. Для оцінки статистики старіння ПЗ/форматів при довготерміновому зберіганні (і відповідних загроз) досліджувалася статистика розвитку проектів розробки ПЗ. З цією метою розглядалися проекти ПЗ з відкритим вихідним кодом, а саме – статистика розподілу часу між публікаціями чергових версій ПЗ або чергових пакетів розширень.

У результаті аналізу дат публікації пакетів розширень із загального мережевого архіву (CRAN) R-мови програмування було побудовано розподіл пакетів, ранжируваних за часом між їх публікаціями. Він може бути апроксимований за допомогою степеневої функції з достовірністю апроксимації майже 0,97, що дозволяє припустити степеневий характер статистики старіння ПЗ (Рис. 3). При представленні отриманого розподілу у подвійній логарифмічній шкалі графік приблизно відповідає прямій лінії, що підтверджує наявність степеневого закону (Рис. 4).

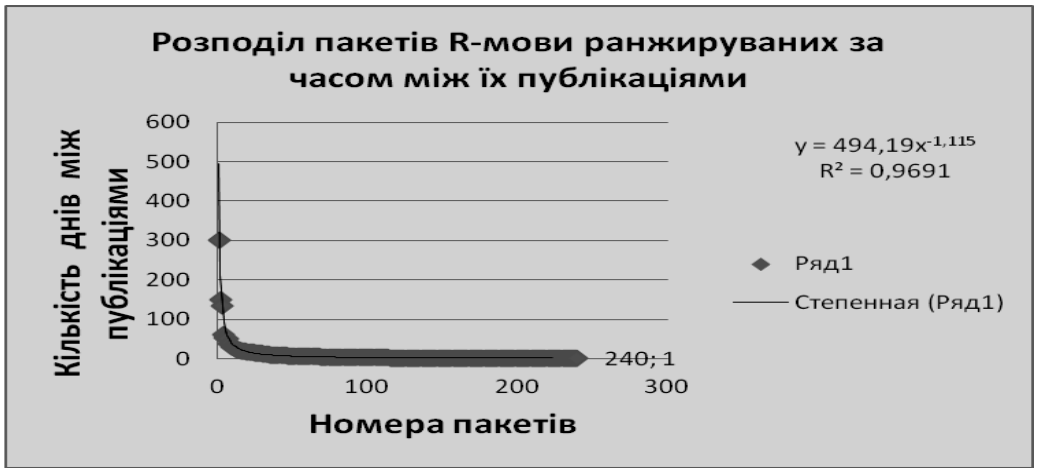

Рис. 3. Розподіл пакетів R-мови, ранжируваних за часом між їх публікаціями з апроксимацією степеневою функцією.

Рис. 4. Розподіл пакетів R-мови, ранжируваних за часом між їх публікаціями у подвійній логарифмічній шкалі.

Аналіз статистики про інші проекти відкритого ПЗ (GCC – набір компіляторів, Ruby – мова програмування) показав більший коефіцієнт достовірності при апроксимації експоненціальною функцією, що може пояснюватися недостатнім обсягом зібраної статистики.

В роботі [12], з метою дослідження старіння форматів даних, аналізувалися веб-ресурси, що використовувалися в домені UK (Великобританія). За допомогою програмних засобів DROID та Apache Tika, а також ресурсів Internet Archive було проаналізовано біля 2,5 млн. веб-ресурсів за період 1996 – 2010 рр. Отримані результати наведено на Рис. 5 та Рис. 7. На цих рисунках по осі абсцис показано роки, протягом яких спостерігалися зміни версій форматів HTML та PDF, а по осі ординат – відсотки веб-ресурсів, на яких використовувались відповідні версії форматів.

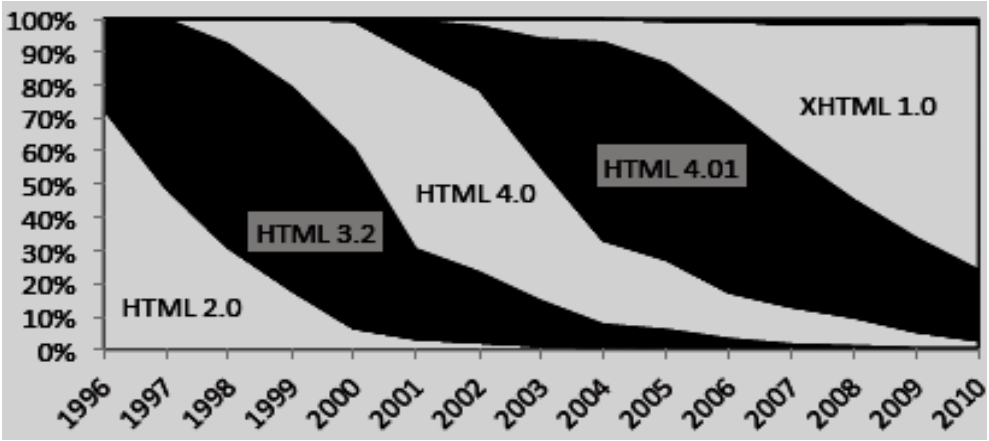

Рис. 5. Зміна версій формату HTML, який використовувався у веб-ресурсах домену UK в 1996 – 2010 рр.

Для отримання статистики старіння форматів та використання її в розглянутих вище імітаційних моделях, було проведено оцінку розподілу частки використання різних версій форматів HTML та PDF у 1996 – 2010 рр.

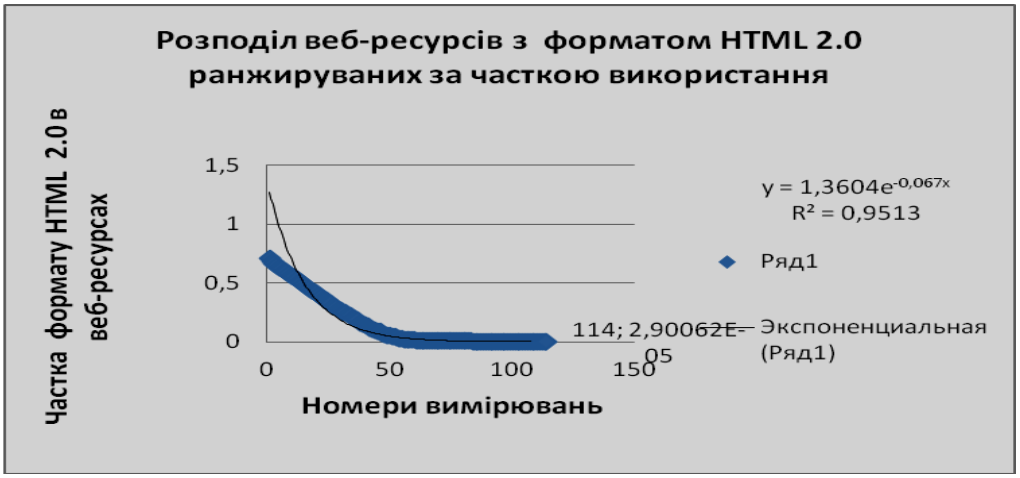

Рис. 6. Дані про розподіл веб-ресурсів з форматом HTML 2.0, ранжируваних за часткою використання в домені UK у 1996 – 2010 рр. з апроксимацією експоненціальною функцією.

З цією метою, результати, наведені на Рис. 5 та Рис. 7, було проранжировано та апроксимовано. На Рис. 6 та Рис. 8 показано, що для формату HTML 2.0, отриманий розподіл апроксимується експоненціальною функцією з достовірністю, більшою ніж 0,9, а для формату PDF 1.2 – розподіл апроксимується степеневою функцією з достовірністю, близькою до 0,9.

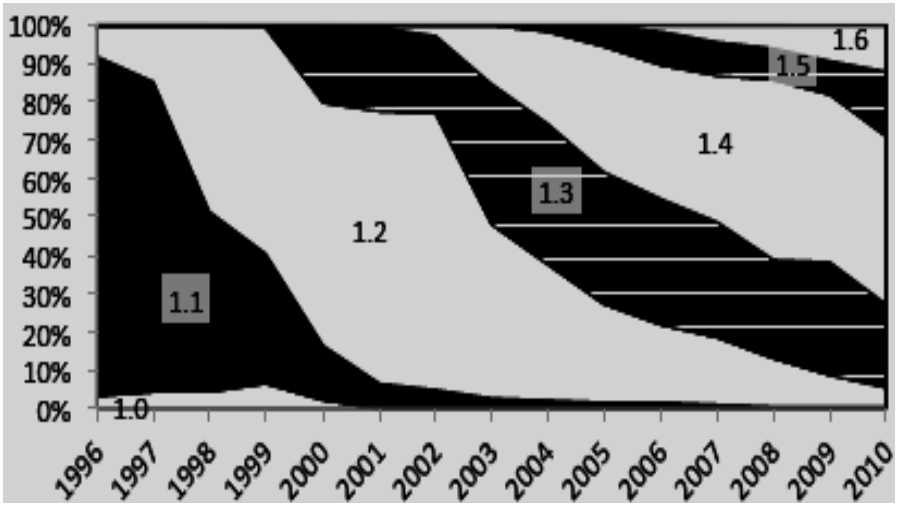

Рис. 7. Зміна версій формату PDF, який використовувався у веб-ресурсах домену UK в 1996 – 2010 рр.

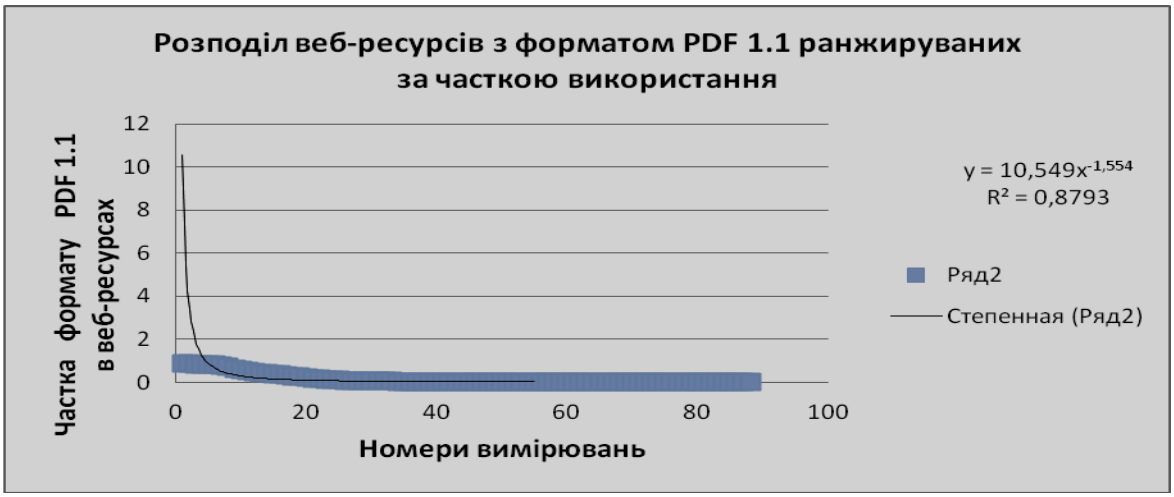

Рис. 8. Дані про розподіл веб-ресурсів з форматом PDF 1.2, ранжируваних за часткою використання в домені UK у 1996 – 2010 рр. з апроксимацією степеневою функцією.

Аналогічно показаному на Рис. 6 та Рис. 8, було проранжировано та апроксимовано дані для інших версій форматів з Рис. 5 та Рис. 7. Ці результати використовуються при створенні імітаційної моделі, яка серед інших враховує також і загрози старіння форматів. Актуальність врахування старіння форматів на основі аналізу зміни їх версій на веб-ресурсах підтверджується створенням сервісу постійних посилань за допомогою зберігання архівних копій веб-ресурсів [13]. (Новий сервіс Perma.cc спрямовано на вирішення проблеми забезпечення стабільності URL та довготермінового зберігання веб-ресурсів).

Моніторинг загроз. Забезпечення живучості ІО в умовах негативних впливів передбачає оцінку цих впливів з метою вибору адекватної реакції, відмови від деяких функцій. Тобто однією з важливих задач забезпечення живучості є моніторинг негативних впливів [14, 15], загроз довготерміновому зберіганню. Моніторинг щодо виявлення загроз повинен здійснюватися з використанням запропонованих моделей загроз. Як зазначалося вище, планування зберігання відповідно до моделі OAIS теж передбачає моніторинг змін у технологічному середовищі та у вимогах користувачів як одну з основних функцій.

У роботі [16] наголошується, що в довготерміновому зберіганні моніторинг є ключовою функцією, яка забезпечує раннє виявлення загроз. Проте, оскільки обсяг та різноманітність загроз зростають, стає неможливим ручний моніторинг усіх аспектів середовища, які можуть заважати зберіганню. Більше того, моніторинг повинен виявляти не тільки ризики зберігання, а й сприятливі можливості (наприклад, зменшення витрат) та гарантувати, що дії по зберіганню, визначені процесами керування, досягають цілей та виправдовують сподівання.

Оскільки запропоновані вище моделі теж направлені на виявлення загроз, то розробка таких моделей може розглядатися як складова частина моніторингу загроз. Тобто не тільки виявлення загроз, а й удосконалення, оновлення та розробка нових моделей загроз повинні здійснюватися протягом всього життєвого циклу довготермінового зберігання. Таку діяльність доцільно організовувати на основі мережі відповідних центрів компетенції, які обмінюються інформацією між собою.

Висновки.

На базі значного статистичного матеріалу здійснено моделювання основних видів загроз, негативних впливів при довготерміновому зберіганні великих обсягів даних. Показано важливе місце степеневого розподілу в цих моделях.

Побудовані моделі, особливості статистики їх розподілів є основою розробки методів планування довготермінового зберігання для забезпечення живучості інформаційних об’єктів.

Оновлення, удосконалення моделей загроз (як і безпосередньо моніторинг загроз) повинні здійснюватися протягом всього життєвого циклу зберігання на основі відповідних центрів компетенції.

Використана література

1. About LexisNexis. – Режим доступу : //www.lexisnexis.com/en-us/about-us/about-us.page

2. США : основные положения Типового закона о правовых актах, публикуемых в электронном виде. – Режим доступу : //www.rusrim.blogspot.com/2013/04/blog-post_7303.html

3. Rodes S., Neacsu D. Preserving and ensuring long-term access to digitally born legal information // Information and Communication Technology Law – 2009. – Vol. 18. – No.1. – P. 39-74.

4. Rhodes S. Breaking Down Link Rot: The Chesapeake Project Legal Information Archive’s Examination of URL Stability // Law Library Journal. – 2010. – Vol. 102 – No. 33. – P. 581-597.

5. Danner R. A., Leong K., Miller W. V. The Durham Statement Two Years Later: Open Access in the Law School Journal Environment // Law Library Journal. – 2011. – Vol. 103. – No. 2. – P. 39-54.

6. Luan F., Nygård M., Mestl T. A Mathematical Framework for Modeling and Analyzing Migration Time // Proceedings of the 10th annual joint conference on Digital libraries , JCDL’10, 2010. – P. 323-332.

7. Weihs C., Rauber A. Simulating the Effect of Preservation Actions on Repository Evolution. “Proceedings of the 8th International Conference on Preservation of Digital Objects”, herausgegeben von: National Library Board Singapore; Nanyang Technical University Singapore. – Singapore, 2011. – P. 62-69.

8. Ланде Д.В., Березін Б.О. Живучість інформаційних об’єктів при довготерміновому зберіганні великих об’ємів даних // Інформація та безпека. – 2012. – № 3-4 (11-12). – С. 13-15.

9. Березін Б.О., Ланде Д.В. Оцінка живучості інформаційних об’єктів при довготерміновому зберіганні великих обсягів даних : *материалы международной научной конференции ИТБ-2013 [“Информационные технологии и безопасность. Оценка состояния”]* : – К. : ИПРИ НАН Украины, 2013. – С. 21-27.

10. Додонов А.Г., Ландэ Д.В. Живучесть информационных систем. – К. : Наук. думка, 2011. – 256 с.

11. Березін Б., Ланде Д. Дослідження стану оптичних носіїв при довготерміновому зберіганні цифрової інформації // Студії з архів. справи та документознавства. – 2012. – Т. 20. – С. 133-139.

12. A. N. Jackson. Formats over time: Exploring uk web history. Proceedings of the 9th International Conference on Preservation of Digital Objects, October 2012. – P. 155-158.

13. About Perma. – Режим доступу : //www.perma.cc/about

14. Додонов А.Г., Флейтман Д.В. Технологические аспекты обеспечения живучести информационных систем // Известия Таганрогского государственного университета. – 2005. – Т. 48. – № 4. – С. 5-7.

15. Бойченко А.В. Вимоги до систем моніторингу факторів впливу на живучість // Реєстрація, зберігання і обробка даних. – 2008. – № 1. – С. 103-115.

16. Faria L., Petrov P., Duretec K., Becker C., Ferreira M., Ramalho J. Design and architecture of a novel preservation watch system // In International Conference on Asia-Pacific Digital Libraries, 2012. – P. 168-178.

~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~
~~~~~~~~~~~~~