# ДИГИТАЛИЗАЦИЯ!? WEB АРХИВИРАНЕ!? ПРОБЛЕМ И/ИЛИ РЕШЕНИЕ

***Ваня Илиева***
*Нов български университет, гр.София*
*vruseva@nbu.bg*

# DIGITALIZATION AND WEB ARCHIVING. PROBLEMS AND SOLUTIONS

***Vanya Ilieva***
*New Bulgarian University , Sofia*

**ABSTRACT:** In relation to the presentations made on LIBER 37th General Conference, the author would like to present a short introduction to the digitalization and web archiving activities. Presented are generalizations and possible solutions of the problems derived from the case studies of various leading libraries, projects and experience in the digitalization and web archiving.

Домакин на 37-мата годишна конференция на LIBER, проведена от 1-ви до 5-ти юли 2008 г. в един от най-древните, натежали от история, и красиви, съвременни градове Истанбул, бяха Koç University и неговата библиотека Suna Kıraç. Добрата организация, интересните презентации и дискусии, социалните мероприятия и фирменото изложение на водещи издателства и агрегатори на DB, присъщи за всички конференции на LIBER, и тази година бяха перфектни.
Част от заглавията на презентации, изнесените в унисон с основната тема на конференцията, предоставящи ценна информация са:

- "Web Archiving in the UK: Co-operation, Legislation and Regulation", John Tuck, Head of British Collections
- "Re-Inventing Collection Development Policy in the Age of Web Archiving : Experience of the National Library of France", Gildas Illien, Project manager for Digital legal deposit, BnF
- Webarchiving: Issues and Problems in Collection Building and especially Access by Grethe Jacobsen The Royal Library (National Library of Denmark)
- Library and Publisher Collaboration on a Graduate Student Portal: A Public Information Service to Enhance Student Productivity Austin McLean Director, Scholarly Communication
- DRIVER: Building a sustainable infrastructure of (European) Scientific Repositories Norbert Lossau, Scientific Coordinator Göttingen State and University Library

Споделения чрез тях опит в дигитализацията и уеб архивирането, натрупан и развиван от водещи световни библиотеки, представените софтуерни разработки,

подобряващи дейността на библиотеките, подпомагат очертаването на стратегии за развитие и постигане на усъвършенстване в предоставянето на информация и обслужване на потребителите и в българските библиотеки.
Сайтът на 37 годишна конференция на LIBER на интернет адрес:
http://www.ku.edu.tr/ku/index.php?option=com_content&task=view&id=2774&Itemid=3340
предоставя възможност всеки заинтересован да разгледа, да изтегли публикуваните там презентации, както и богат архив от снимки, направени по време на конференцията и на организираните социални мероприятия.
С примери от презентациите ще ви запозная с тенденциите, проблемите и опита на колеги от водещи библиотеки и сродни информационни организации по отношение на:

- Дигитализация: Проблеми & решения
- Web архивиране: Проблеми & решения

Дигитализация е процесът на конвертиране на информация в цифров формат. С развитието на технологиите тя все повече се превръща в предпочитан метод за разрешаване на проблеми, които не засягат само нашата библиотечна общност, но и тези в развити страни като Великобритания, Франция, Холандия, Дания. Най-общо казано дигитализацията е поредица от дейности, насочени към запазване, дългосрочно съхранение на материали, значително намаляване на размерите на пространството за физическо съхранение, спестяване на дублирането на едни и същи документи в различни библиотеки, спестяване на финансови средства и разхищение на висококвалифициран труд, осигуряване на лесен, бърз и отдалечен достъп до тях. Всичко това в крайна сметка е в полза за решението и на основния проблем – предлагане на адекватно обслужване на специалисти, изследователи, научни работници, студени, читатели от всякакъв ранг в динамичното, забързано високотехнологично време.
Изброените проблеми се решават в сътрудничество, в партньорство между библиотеките от една страна, и от друга между тях и издателства, държавни и международни институции, занимаващи се с авторско право, фирми – разработчици на софтуер, организации от всякакъв тип, предлагащи финансова помощ. Изтъквам това, защото все още в България нивото на кооперативност не е на такава висота. Все още библиотеките самостоятелно, а не в обединение дигитализират части или цели колекции. В презентацията си за развитието на проекта Europeana Пол Айрис казва “библиотеките се насочват от единични дигитализационни проекти към мащабни такива, които в един момент ще доведат до съществуването на цели колекции с милиони книги. Инвестиционните агенции подпомагат изследователските и демострационните проекти, които помагат на библиотеките и институциите, управляващи културното наследство да

постигнат по-добро разбиране на процеса на дигитализация ” Т.е. само в партньорство би могло да се постигне максимализъм в процеса на дигитализация с крайните ясно очертани цели и резултати. Във всички посочени презентации ще срещнем дебело подчертаната фраза: Необходимост от добре организирано партньорство!

Разрешавайки тези съществени проблеми за библиотеките и посочването на изграждане на различни типове обединения около осъществяването на проекти по дигитализация, тя сама по себе си довежда други.

Това са обобщените проблеми, породени от процеса по цифровизиране на информацията:

- Критерий за избора и необходимост от предоставяне на пълно покритие на колекциите
- Необходимост от стандарт за метаданните
- Технически стандарти - за цифровизация, каталогизация и идентификация на документи чрез регистрация или постоянни профили и т.н.
- Авторски права и интелектуална собственост
- Проблеми при търсенето и предоставянето, Скорост при отваряне на файлове
- OAI съвместимост
- Бизнес модели: Свободен достъп/Публични и частни партньорства, права на достъп, Ценообразуване
- Координиране на търгове за финансиране на цифровизиращите дейности
- Подготовка за съхранение на цифровата информация

Изброявам ги не защото не са известни, а за да подчертая, че и колегите ни от водещи библиотеки като националната библиотека на Франция, Британската библиотека, Кралската библиотека на Дания се сблъскват непрекъснато с тях при изграждането на собствените им дигитални колекции и всички те единодушно са констатирали, че партньорството по съгласуване на методологията при цифровизирането е начинът за разрешаване на проблемите.

Отново давам пример от презентацията на Пол Айрис, който в заключение призовава ЛИБЕР като водещо обединение на европейските академични библиотеки, да бъде посредник в Европа със следните функции:

- Изгражда лоби в Европейската комисия, за включване на библиотеките в процеса
- Европейските изследователски библиотеки поставят фокуса на цифровите си стратегии върху крайния потребител
- Изгражда европейски план за цифровизация
- Необходимост от европейски портал за обединяване на отделните ресурси

Уеб архивирането бе друга водеща тема на годишната конференция на LIBER.

Появата на World Wide Web като глобален информационния ресурс е създал специални предизвикателства пред библиотеките за архивиране на цифрови

материали, налични и достъпни в мрежата. Единиците, в които тези материали се измерват са познатите ни web page и web site. Преди да ви запозная конкретно с представените ни по време на конференцията стратегии и изградени вече web архиви, бих искала да направя кратко въведение в web архивирането. То не е непознато, неизвестно, защото предполагам че всички търсят новостите в професионален аспект и поне от любопитство всеки се е запознавал с опита на чуждестранни библиотеки, а може би някои вече са опитвали и изграждали собствени такива колекции.

Моето въведение се състои по-скоро в представяне на следните определения и термини:

- **Web архивиране** (web archiving) е процесът на събиране на части от World Wide Web в смисъла на сайтове, изграждане на архиви и осигуряване на достъп до тях.
- Поради необятността на WWW, специалистите, занимаващи се с изграждането на web архиви, обикновено използват web краулери (**web crawler)** за автоматизирано събиране. Това са програми или автоматизирани скриптове, които търсят, прелистват наличното уеб съдържание в строго определен, методичен начин.
- Други термини: web spider, web robot, web scutter, ants, automatic indexers, bots, worms
- Процесът по търсене чрез тези програми се нарича **уеб индексиране** или **spidering.** Много търсачки използват spidering като средство за предоставяне на актуална информация. Уеб краулерите се използват главно, за да се създаде копие на всички посетени страници, с цел по-нататъшна обработка, индексиране на страниците и изтеглянето им. Краулерите също се използват за автоматизирането на задачите по поддържане на уебархив, по проверка на връзки или за проверка на HTML кода, идентифициране на всички хипервръзки в страницата, и добавянето им към списъка на URL адреси за посещение.
- **Метаданни –** автентичност и произход на web ресурсите, критерии за търсене в web архив

Нови длъжности, наложили се в западноевропейските библиотеки при процеса на web архивиране:

- **Web archiving programme manager**
- **Web archiving curator**
- **Head of digital infrastructure**

С цитат от презентацията на Гилда Илен представям какъв е подходът за създаване на тези длъжности, все още непознати за българската библиотечна общност:

**‘you need to get your librarians and IT staff to work hand in hand and invent together new job profiles – digital and web curators’**

За всички е ясно, че е невъзможно да се индексира и поддържа архив на цялата мрежа. Три са основните затруднения при уеб индексирането:

- Голям обем – Общият обем цифрова информация през 2006 г. се изчислява на 161 милиарда exabytes

**Принцип:** *архивират се всички web сайтове със значимо за съответна организация web съдържание*

- Собствен темп на изменение и жизнен цикъл на web страница
- Поговорката гласи: една уебстраница има живота е на муха
- Поколението на динамичните уеб страници

**Извод:** Предвид текущите размери на уеб, дори големи търсачки покрива само част от обществено достъпни интернет. Затова всички са изправени пред въпроса как да се справим с тези проблеми, за да удовлетворим отново пропорционално нарастващите на сайтовете информационни нужди на нашите потребители.

Един от подходите за разрешаване на изброениет по-горе проблеми е съчетаване между приетия в съответната срана Закон за задължителен депозит и възприетите принципи на комплектуване. В презентациите си представители на различни водещи библиотеки изтъкнаха факти, налагащи промени и компромиси както в Закона, така и в комлектуването. Ето няколко примера:

- Дания и Франция – допълнения към ЗЗД, чрез които обекти на задължително депозиране стават всички ресурси, публикувани в домейни .dk и .fr, и всички web материали, отнасящи се до история, общество, култура и наука на дадената страна
- Дания – узаконяване на изградената и поддържана от виртуална депозиторна институция netarkivet.dk (netarchive.dk) – web архив
- Холандия – споразумения за доброволен депозит

Друг начин за максимално постигане на целите при изграждане на онлайн архив от сайтове е чрез избор на един от трите основни подхода при процеса на web архивиране:

- **Domain (географски) подход**
- **Селективен подход**
- **Комбинативен подход**

Селективният модел, в основата на който е подбор на сайтовете според тяхното тематично съдържание, е най-разпространен и предпочитан от водещи западни библиотеки, като например британските. Обединеният им web архив е известен под съкращението *UKWAC* и при изграждането и поддържането му вземат участие различни специалисти и цели звена от изброените библиотеки:

*British Library, National Library of Scotland, National Library of Wales, The National Archives, JISC, and the Wellcome Trust*

Резултатът е:

На посочения по-долу адрес има пълен безплатен достъп до всички уебсайтове на изброените по-горе организации: *www.webarchive.org.uk*
Ето и няколко статистически данни от британския *UKWAC*, доказващи че селективния модел при изграждане на web архив е доста успешен:

- *Общ брои сайтове: 2,769*
- *Общ обем информация: 1.9 terabytes*
- *Брои сайтове по партньори:*
  - *British Library - 1,200*
  - *National Library of Wales - 315*
  - *National Library of Scotland - 110*
  - *The National Archives - 371*
  - *JISC - 460*
  - *Wellcome Trust - 324*

Освен подход при изграждане на web архив, е важно да се съблюдават и следните критерии при избор на сайтове, преди включването им в тази своеобразна колекция:

- Времеви обхват
- Език / езици
- Авторски права и интелектуална собственост

*'(almost)* ***anyone can publish*** *(almost)* ***anything*** *(almost)* ***anywhere****'*

- Аудитория
- Формат на web документа: blog, wiki, e-books, news…
- Тип публикация: *правителствени документи, научни сайтове и сайтове на лаборатории, фирмени сайтове…*
- Достъп и защита на лични данни
- Бюджет
- Персонал

В заключение, искам да представя някои фактологични данни, с които да докажа че не отдавна web архивирането се налага като тенденция в библиотечната общност, но въпреки това има утвърден опит и успешно реализирани проекти по създаване на web архиви, формирайки стабилна основа, на която да стъпим и да поставим начало на създаване и на български web архив:

**Факти & статистика**

- 1996 г. - Web архивиране
- Internet Archive
- Национална библиотека на Австралия – селективен модел
- Национална библиотека на Швеция – domain (географски) модел
- 1997 г. кооперирание в web архивирането на международно ниво
- 2003 г. International Internet Preservation Consorcium
- 15 европейски национални библиотеки web архивират
- 9 европейски национални библиотеки са процес на планиране и тестване.