

**Národní knihovna České republiky**

**Registrace, ochrana a zpřístupnění domácích  
elektronických zdrojů v síti Internet**

**Závěrečná zpráva za léta 2000-2001**

**Předkládá: PhDr. Vojtěch Balík, ředitel NK ČR**

**Zpracovala: Mgr. Ludmila Celbová, řešitelka**

**Praha, leden 2002**

# OBSAH

A	KONSTATAČNÍ ČÁST .....	3
A.1	<b>Rešerše .....</b>	<b>3</b>
A.1.1	Publikační činnost k řešení projektu.....	3
A.1.2	Odkazy na významné informační zdroje použité k řešení projektu.....	3
A.2	<b>Současný stav ve světě a v ČR .....</b>	<b>4</b>
A.3	<b>Cíl.....</b>	<b>5</b>
B	ANALYTICKÁ ČÁST .....	6
B.1	<b>Vlastní řešení .....</b>	<b>6</b>
B.2	<b>Přínos řešitele.....</b>	<b>7</b>
B.2.1	Oblast problematiky vztahů knihoven, vydavatelů a legislativy .....	7
B.2.2	Oblast problematiky informačních a komunikačních technologií .....	11
B.3	<b>Posun znalostí .....</b>	<b>14</b>
C	NÁVRHOVÁ ČÁST .....	15
C.1	<b>Výsledky řešení .....</b>	<b>15</b>
C.1.1	Spolupráce s vydavateli.....	15
C.1.2	Legislativa .....	15
C.1.3	Informační a komunikační technologie.....	16
C.2	<b>Propagace projektu .....</b>	<b>17</b>
C.3	<b>Závěr.....</b>	<b>18</b>
C.4	<b>Návrhy opatření .....</b>	<b>18</b>
D	RESUMÉ A KLÍČOVÁ SLOVA .....	19
D.1	<b>Resumé .....</b>	<b>19</b>
D.2	<b>Klíčová slova.....</b>	<b>19</b>
E	PŘÍLOHY .....	20

# A KONSTATAČNÍ ČÁST

## A.1 Rešerše

### A.1.1 Publikační činnost k řešení projektu

1. CELBOVÁ, Ludmila. Informace o projektu registrace domácích internetových zdrojů nově na serveru WebArchiv. *Ikaros* [online]. 2001, č. 5 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://ikaros.ff.cuni.cz/2001/c05/webarchiv.htm>>. ISSN 1212-5075.
2. CELBOVÁ, Ludmila. Katalogizace elektronických zdrojů na Internetu: proč, co, jak?. *Ikaros* [online]. 2001, č. 2 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://ikaros.ff.cuni.cz/2001/c02/katalogizace.htm>>. ISSN 1212-5075.
3. CELBOVÁ, Ludmila. Stanou se online dostupné elektronické zdroje integrovanou součástí digitálních knihoven? *Národní knihovna*, 2001, roč. 12, č. 2, s. 91-98. ISSN 0862-7487.
4. CELBOVÁ, Ludmila. WebArchiv: Projekt zaměřený na českou národní bibliografii elektronických zdrojů. *ITlib*, 2001, roč. 5, č. 3, s. 38-41. ISSN 1335-793X.
5. VOJTÁŠEK, Filip. Archiv celosvětového webu zpřístupněn. *Ikaros* [online]. 2001, č. 12 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://ikaros.ff.cuni.cz/2001/c12/archive.htm>>. ISSN 1212-5075.
6. VOJTÁŠEK, Filip; ŽABIČKA, Petr. Workshop mnoho konkrétního z realizace projektu NEDLIB nepřinesl. *Ikaros* [online]. 2001, č. 1 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://ikaros.ff.cuni.cz/2001/c01/nedlib.htm>>. ISSN 1212-5075.

### A.1.2 Odkazy na významné informační zdroje použité k řešení projektu

1. *Biblink* [online]. [Bath (Anglie) : UKOLN], last updated 12-Jul-2000 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://hosted.ukoln.ac.uk/biblink>>.
2. *Cobra+ : Computerised Bibliographic Record Actions : factsheet* [online]. Boston Spa : British Library, [1997] [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://portico.bl.uk/gabriel/en/projects/cobra.html>>.
3. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal of the European Communities : english edition* [online]. 2001, č. L 167 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <[http://europa.eu.int/eur-lex/pri/en/oj/dat/2001/l\\_167/l\\_16720010622en00100019.pdf](http://europa.eu.int/eur-lex/pri/en/oj/dat/2001/l_167/l_16720010622en00100019.pdf)>. ISSN 0378-6978
4. *Dublin Core Metadata Initiative : making it easier to find information* [online]. [Ohio (USA)] : DCMI, [2001-01-16] [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://purl.oclc.org/dc>>.
5. *EVA : the acquisition and archiving of electronic network publications* [online]. Helsinki (Finsko) : Helsinki University Library, updated 15. 12. 1997 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://www.lib.helsinki.fi/eva/english.html>>.
6. *INDOREG : INternet DOcument REGistration : project report* [online]. Ballerup (Dánsko) : Dansk Bibliotheks Center, 16. 9. 1997 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://www.purl.dk/rapport/html.uk>>.

7. International declaration on the deposit of electronic publications : Conference of European National Librarians/Federation of European Publishers (CENL/FEP). *Dialog mit Bibliotheken*, 2000, vol. 12, no. 3, s. 2-14. ISSN 0936-1138. Dostupné též na World Wide Web: <<http://www.ddb.de/news/epubstat.htm>>.
8. *Kulturarw<sup>3</sup> Heritage Project* [online]. Stockholm (Švédsko) : Royal Library, [1998] [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://kulturarw.kb.se/html/projectdescription.html>>.
9. *LC21: a Digital Strategy for the Library of Congress (2001)* [online]. Washington (USA) : National Academy Press, c2001 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://bob.nap.edu/books/0309071445/html>>.
10. *Librarians and publishers working to a common agenda* [online]. Hague (Nizozemí) : IFLA, c1995-2000, latest revision August 31, 2001 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://www.ifla.org/III/misc/pr310801.htm>>.
11. *NEDLIB : Networked European Deposit Library* [online]. Hague (Nizozemí) : Koninklijke Bibliotheek, c1998, last updated 11 March 2001 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://www.kb.nl/nedlib>>.
12. *Networked Electronic Publications Policy and Guidelines : Electronic Collections Coordinating Group, National Library of Canada, October 1998* [online]. Ottawa (Kanada) : National Library of Canada, last updated 2001-04-01 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://www.nlc-bnc.ca/publications/8/index-e.html>>.
13. *Nordic Web Archive* [online]. [Oslo (Norsko) : National Library of Norway, 2000] [cit. 19. 7. 2001]. Dostupné na World Wide Web: <<http://nwa.nb.no>>.
14. *OCLC CORC* [online]. Dublin (Ohio, USA) : OCLC, c2001 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://www.oclc.org/corc>>.
15. *PANDORA Archive : Preserving and Accesing Networked Documentary Resources of Australia* [online]. [Canberra (Austrálie)] : National Library of Australia, [2001] [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://pandora.nla.gov.au/index.html>>.
16. *RLG and OCLC Explore Digital Archiving* [online]. [Mountain View (USA) : Research Libraries Group, 2000-03-10] [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://www.rlg.org/pr/pr2000-oclc.html>>.
17. *The Nordic Metadata projects* [online]. Helsinki (Finsko) : Helsinki University Library, [5. 3. 1999], last updated 21 February 2000 [cit. 22. 1. 2002]. Dostupné na World Wide Web: <<http://www.lib.helsinki.fi/meta>>.

## A.2 Současný stav ve světě a v ČR

Během dvouletého řešení projektu v Národní knihovně ČR se nejen značně rozšířil počet zemí, v nichž se začali systematicky zabývat problematikou registrace, archivace a trvalého zpřístupňování elektronických zdrojů publikovaných v síti Internet, ale tato komplexní problematika se také dostala do čela okruhů problémů řešených na poli mezinárodních institucí – zejména společná iniciativa CENL (Conference of European National Librarians) a FEP (Federation of European Publishers), která byla řešiteli ve stávajícím projektu zohledněna, nebo nejnověji společná iniciativa IFLA (International Federation of Library Associations and Institutions) a IPA (International Publishers' Association). Problém trvalého uchování národního bohatství v podobě elektronických publikací, zejména síťových, tedy už přestává být experimentem „pokrokovějších“ zemí, ale stává se obecně naléhavou výzvou pro knihovny i nakladatele, neboť objem

digitálních informací narůstá obrovským tempem a politice ochrany těchto dokumentů a k tomu sloužícím technologiím se dosud věnovala minimální nebo nulová podpora. Podle společného vyjádření IFLA a IPA datovaného 14. ledna 2002 je potřeba archivace velmi naléhavá – odhaduje se, že mnohé z elektronických zdrojů „vzniklých jako digitální“, tedy zdroje, které nemají souběžnou kopii v jiné (nejčastěji tištěné) formě, byly již trvale ztraceny, neboť jejich tvůrci odstranili z webu své elektronické publikace, aniž by zajistili jejich trvalou archivaci. „I když náklady na dlouhodobou archivaci jsou vysoké, náklady na nicedělání v této oblasti by byly katastrofální.“

Ukazuje se, že v České republice se začalo s řešením komplexní problematiky registrace, ochrany a zpřístupňování elektronických publikací právě včas. V rámci dvouletého pilotního projektu byla provedena nejprve analýza řešení problematiky ve světě na národní i mezinárodní úrovni a na jejím základě se řešitelé orientovali zejména na řešení v evropských severovýchodních zemích, kde jsou výsledky jak v oblasti knihovnické a legislativní, tak i v oblasti technické velmi progresivní a zároveň jsou kulturně-sociální podmínky i technologické podmínky (rozsah a struktura „domácího“ webu) srovnatelné se stavem v ČR.

Po celou dobu řešení pilotního projektu, který dostal pracovní název WebArchiv, spolupracovala řešitelská instituce, Národní knihovna České republiky, s pracovníky Ústavu výpočetní techniky Masarykovy univerzity v Brně v oblasti problematiky informačních a komunikačních technologií; na řešení okruhu problémů knihovnických a legislativních se podíleli externí spolupracovníci – odborníci v oblasti elektronického publikování (časopis Ikaros - Ikaros, o. s.).

### **A.3 Cíl**

Cílem projektu bylo připravit podmínky pro zpracování české národní bibliografie elektronických zdrojů, se zaměřením zejména na zdroje dálkově přístupné. S bibliografickým zpracováním souvisí zajištění trvalého uchování monografických i seriálových domácích elektronických dokumentů publikovaných v síti Internet a jejich zpřístupnění, respektující autorské právo vydavatelů.

Komplexní problematika registrace, ochrany a zpřístupnění elektronických síťových zdrojů zahrnuje oblast problémů knihovnických, legislativních a problémů z oblastí informačních a komunikačních technologií, které se navzájem prolínají. V oblasti knihovnické bylo třeba vytvořit metodiku pro výběr dokumentů a jejich zpracování s aplikací národních a mezinárodních standardů. Legislativní otázky, jimiž se řešitelé zabývali, se týkají povinného výtisku, resp. prozatímního řešení otázky oprávnění k získávání, archivaci a zpřístupňování elektronických zdrojů, a dále řešení autorskoprávní problematiky, tedy využívání archivovaných a zpřístupňovaných elektronických zdrojů. V oblasti ICT bylo třeba vyvinout softwarové nástroje umožňující získávání, archivaci, zpracování a zpřístupňování elektronických zdrojů při dodržení legislativních podmínek.

## B ANALYTICKÁ ČÁST

### B.1 Vlastní řešení

Problematika řešená v tomto projektu je velmi komplexní, zahrnuje oblast knihovnictví a vydavatelství, práva i informačních technologií. Navíc vyžaduje aplikaci mezinárodních standardů a kompatibilitu řešení s jinými podobnými projekty zpracovávanými ve světě v nedávné minulosti i v současnosti, na nichž pracují velké týmy specialistů.

Řešitelé proto věnovali v prvním roce práce na projektu značnou pozornost informačním průzkumům, získání dostupných informačních materiálů publikovaných v tištěné a zejména v elektronické formě, jejich analýze a navázání kontaktů s vytypovanými zahraničními pracovišti, od nichž lze získat cenné informace, zkušenosti i softwarové nástroje jako výsledky řešených národních i mezinárodních projektů. Tento postup je racionální a ekonomicky únosný v rámci finančních, personálních a organizačních možností českého knihovnictví a současně by mohl umožnit zařadit se poměrně rychle mezi pokrokové země v oblasti získávání, ochrany a zpřístupňování elektronických dokumentů publikovaných v síti Internet.

Jedním z nejrozsáhlejších a nejvýznamnějších zahraničních projektů na tomto poli byl projekt NEDLIB, ukončený v roce 2000, na jehož realizaci se podílelo osm národních knihoven západoevropských zemí a tři instituce zajišťující technickou stránku řešení. Tento projekt, který navázal na řadu podobně zaměřených národních i nadnárodních projektů (zejména projekty Nordic Metadata I, II a Nordic Web Archive severovýchodních zemí) a jehož předmětem bylo budování depozitní knihovny elektronických dokumentů, se zabýval všemi elektronickými dokumenty, včetně méně problematických off-line zdrojů. Díky jeho širokému záběru bylo vhodné a možné převzít mnohé z toho, čeho v něm bylo dosaženo. Velkým přínosem pro projekt bylo zejména využití výsledků v oblasti ICT.

Možnost získání zkušeností z řešení klíčových projektů při zahraniční cestě do Finska (Helsinki University Library) na podzim roku 2000 napomohla do značné míry k technickému zaměření řešení na využití softwarových nástrojů, které jsou výsledkem výše zmíněných projektů a které bylo možné získat za výhodných podmínek. K účelům testování těchto nástrojů na vybraném vzorku elektronických zdrojů byly pořízeny dva terminály s odpovídajícími technickými parametry a jeden vysoce výkonný PC s relativně velkými kapacitami operační i diskové paměti pro připojení na síť, pracující v OS Linux. Tento stroj ve fázi testování suploval unixový server, sloužil k instalování nástrojů pro stahování dokumentů, pro ukládání údajů pro popis zdrojů aj. a pro ukládání zdrojů do webového archivu. Na tomto serveru byla také zřízena webová prezentace projektu (na adrese <http://webarchiv.nkp.cz>). V rámci této prezentace byly postupně shromážděny veškeré informace o řešeném projektu.

## B.2 Přínos řešitele

Jak bylo uvedeno již v souhrnné zprávě o projektu za rok 2000, řešení pilotního projektu představuje principiálně testování dvou metod, které by v optimálním případě měly být aplikovány paralelně s cílem umožnit dlouhodobé uchování a využívání elektronických zdrojů:

- shromažďování, registrace a archivace vybraných domácích elektronických online dostupných dokumentů jako legitimní součásti národní publikační produkce podle stanovených kritérií výběru pro účely České národní bibliografie; tato činnost klade značné nároky na intelektuální práci zpracovatelů;
- shromažďování a archivace domácích zdrojů z Internetu v relativní úplnosti (automatizovaný proces).

Ve druhém roce řešení projektu bylo započato s testováním obou metod.

Informace o řešení jsou dále rozděleny do dvou částí – 1) na oblast problematiky knihovnické, resp. vydavatelské a právní a 2) na oblast problematiky informačních technologií.

### B.2.1 Oblast problematiky vztahů knihoven, vydavatelů a legislativy

#### a) *Legislativa*

Mnoho pozornosti bylo věnováno legislativním otázkám, kde došlo v zahraničí ke značnému posunu. Některé země (Nizozemí, Velká Británie, Německo) aplikovaly dohody mezi depozitními knihovnami a vydavateli, umožňující dočasné řešení legislativních otázek získávání, archivace a zpřístupňování zdrojů publikovaných v síťovém prostředí. Velkým přínosem pro řešení této problematiky v dalších zemích by měla být spolupráce mezi CENL (Conference of European National Librarians) a FEP (Federation of European Publishers) – viz dále odst. C.1.2 Legislativa.

#### b) *Využití metadat*

Jelikož metadata hrají při dlouhodobé archivaci a zpřístupnění webových zdrojů zásadní roli, byla této otázce věnována značná pozornost také při řešení projektu. Pro tyto účely byl zvolen Dublin Core Metadata Element Set (zkráceně označovaný Dublin Core), který je všeobecně považován za nejperspektivnější metadatový formát, i když je rozšířen méně, než se původně předpokládalo. K hlavním přednostem Dublin Core patří rozšiřitelnost a opakovatelnost, což znamená, že základní sadu prvků a jejich atributů (kvalifikátorů) je možné doplnit nebo upravit tak, aby vyhovovala potřebám využití tohoto formátu v konkrétních podmínkách.

Využití Dublin Core v projektu se opírá o překlad Dublin Core Metadata Element Set (v poslední verzi 1.1 schválené v červenci 1999) a kvalifikátorů (schválené v dubnu 2000) do češtiny. Tento překlad byl publikován v červnu 2000 (dostupný na

[http://www.ics.muni.cz/dublin\\_core/DC-czech-1.1.html](http://www.ics.muni.cz/dublin_core/DC-czech-1.1.html)); jeho garantem je Ústav výpočetní techniky MU.

Nejzásadnější změna Dublin Core, která byla provedena v rámci projektu WebArchiv a která se promítla do lokalizované verze generátoru metadatových záznamů podle tohoto standardu, se týkala prvku Předmět (Subject). Původní tzv. kvalifikátory schématu umožňující věcný popis pomocí v zahraničí používaných řízených předmětových slovníků a systematických třídění byly nahrazeny, resp. doplněny slovníky a tříděními používanými v domácích veřejných a odborných knihovnách:

- Předmětová hesla Národní knihovny (ukázkový soubor je publikován na serveru NK)<sup>1</sup>
- Polytematický strukturovaný heslář (spravován Státní technickou knihovnou)<sup>2</sup>
- tezaurus Agrovoc (česká verze spravována Ústavem zemědělských a potravinářských informací)
- tezaurus Eurovoc (česká verze spravována Parlamentní knihovnou ČR)<sup>3</sup>
- Český teologický tezaurus (spravován knihovnou Evangelické teologické fakulty UK)<sup>4</sup>
- tezaurus MeSH (česká verze spravována Národní lékařskou knihovnou)<sup>5</sup>
- MDT-Master Reference File (přístupný online a na CD-ROM)<sup>6</sup>
- MDT-Vybrané znaky<sup>7</sup>

### c) Kritéria výběru

Při formulaci kritérií, podle nichž budou vybírány ty zdroje, které budou zařazovány do České národní bibliografie, se vycházelo ze strategií archivace webových zdrojů přijatých v rámci obdobných zahraničních projektů (zejména projektu National Library of Australia PANDORA), ovšem s přihlédnutím ke specifické situaci v České republice. Kritéria jsou výsledkem diskusí, které provázely experimentální zpracování webových zdrojů podle Dublin Core (ve spolupráci s kolegy z odborných knihoven a vydavateli elektronických seriálů) v druhé fázi řešení projektu. Na tomto místě je třeba zdůraznit, že i nadále platí, že míra úspěšnosti projektů, které budou zaměřeny na rutinní zpracování webových zdrojů, bude odvozena od ochoty vydavatelů těchto zdrojů integrovat Dublin Core do svých publikačních aktivit.

---

<sup>1</sup> [http://www.nkp.cz/standard/Veczprac/vec\\_zprac.htm](http://www.nkp.cz/standard/Veczprac/vec_zprac.htm)

<sup>2</sup> <http://www.stk.cz/cgi-bin/dflex/CZE/STK/PSH>

<sup>3</sup> [http://www.psp.cz/kps/knih/e\\_mtabc.htm](http://www.psp.cz/kps/knih/e_mtabc.htm)

<sup>4</sup> <http://www.etf.cuni.cz/~library/zdroje/cttweb/index.htm>

<sup>5</sup> <http://194.108.164.2:4001/ALEPH/CZE/NML/CLA/MES/SCAN>

<sup>6</sup> <http://aip.nkp.cz/mdt/>

<sup>7</sup> <http://www.nkp.cz/standard/vybrzn/PREHLTRA.htm>

## Kritéria:

### 1. podle domény (místa uložení zdroje)

Primárně jsou brány v úvahu zdroje přístupné na serverech s doménou prvního stupně .cz. V této souvislosti však vyvstává problém, jak správně vymezit tzv. národní web (tj. zda uplatňovat pouze teritoriální hledisko nebo také jazykové hledisko podobně, jako je tomu u tradičních bohemikálních dokumentů). Faktem zůstává, že není možné výše uvedenou podmínku za všech okolností striktně dodržet, protože v některých případech čeští vydavatelé záměrně nebo nuceně (obvykle z důvodu předchozí registrace žádané domény ze strany spekulantů) využívají servery s doménami .com, .net a výjimečně také .org (např. oficiální prezentace Městského úřadu v Hořicích na [www.horice.org](http://www.horice.org), fotografický průvodce New Yorkem od M. Baňkové na [www.nycmap.com](http://www.nycmap.com) aj.). V těchto případech je třeba identifikovat vlastníka domény druhého stupně pomocí specializovaných služeb.<sup>8</sup> Stejně zkušenosti byly získány při automatickém sběru švédských webových zdrojů v rámci projektu Kulturarw<sup>3</sup> – bylo zjištěno, že až 40 % zdrojů je uloženo na serverech mimo národní doménu .se.

### 2. podle obsahu zdroje

Jsou brány v úvahu zdroje odborného nebo uměleckého charakteru, u nichž se předpokládá, že mají informační hodnotu pro větší okruh budoucích uživatelů. Záměrně jsou pominuty zdroje, které jsou výsledkem soukromých, firemních nebo ryze reklamních publikačních aktivit, i když s vědomím, že i v této oblasti se mohou vyskytovat zdroje, které mohou být pro některé uživatele zajímavé, resp. zdroje, které nejsou jiným způsobem zveřejněny.

### 3. podle typu zdroje

Repertoár typů zdrojů (vzorek viz <http://webarchiv.nkp.cz/dc.php>) je poměrně pestrý a je do jisté míry ovlivněn předchozím kritériem. Při jejich výběru se vychází z běžných klasifikací dokumentů. Jde především o seriály, konferenční příspěvky, výzkumné a jiné zprávy, studie vzniklé např. jako výstupy vědeckých a výzkumných projektů, akademické práce, dokumenty veřejné správy. Je příznačné, že tyto zdroje spadají do kategorie tzv. šedé literatury.

### 4. podle formy

Jsou brány v úvahu ty zdroje, které jsou publikovány pouze v elektronické formě, aby se zabránilo duplicitě zpracování webových zdrojů a tradičních (tištěných) dokumentů s identickým obsahem.

---

<sup>8</sup> např. <http://www.allwhois.com/>

## 5. podle přístupu

Jsou brány v úvahu pouze ty zdroje, které jsou volně přístupné, to znamená, že nejsou k dispozici v rámci placených informačních služeb.

## 6. podle formátu

Z pragmatických důvodů jsou preferovány formáty, které jsou všeobecně podporovány producenty aplikačního softwaru (zejména webových prohlížečů), nikoliv tedy proprietární formáty, pro jejichž korektní zobrazení je třeba zvláštní aplikační software. K tomu je třeba dodat, že některé z těchto formátů se staly – díky dominantnímu postavení producenta na trhu – standardy elektronického publikování de facto (např. Adobe – *pdf*, Microsoft – *doc*). Archivaci webových zdrojů usnadňuje empiricky dokázaný fakt (harvesting /stahování/ českého, švédského, nizozemského a finského webu), že navzdory velkému množství formátů, se kterými se na webu můžeme setkat, je většina webových zdrojů (85 až 90 procent) uložena v malém počtu formátů (resp. MIME podtypů) – *html/htm* (k tomu připojme *asp* a *php* v případě dynamických webových informačních systémů), *jpeg*, *gif* (pro statickou grafiku) a *txt*. Zastoupení zdrojů ve formátech *pdf*, *doc*, *rtf* a *ps* (PostScript) na českém webu není výrazné (viz kapitola B.2.2 a příloha F.9), ale jejich informační hodnota je obvykle vyšší než u zdrojů ve formátu *html*.

### d) Elektronické seriály

Zvláštní pozornost byla v rámci projektu soustředěna na elektronické seriály (periodika), u nichž lze očekávat serióznější záměry vydavatele ve srovnání s jednotlivými webovými zdroji. K 31. prosinci 2001 České středisko ISSN registrovalo celkem 221 seriálů přístupných online, které mají různé zaměření: odborné, populárně-naučné, zábavní a zpravodajské. Z teritoriálního hlediska mezi nimi figurují celoplošné i lokální seriály. V jednom případě (Carolina) byla přidělena dvě ISSN (pro českou a anglickou verzi). Ne všechny seriály jsou dosud vydávány. Z uvedeného počtu byl k uvedenému datu pouze časopis *Ikaros* systematicky excerpován v celostátním měřítku (v databázi knihovnické literatury KKL Národní knihovny ČR od srpna 1999 a v článkové polytematické databázi ANL, která je součástí České národní bibliografie, od května 1999). S některými vydavateli byla zahájena spolupráce (*Ikaros*, Česká škola, *Elektrorevue*). Vydavatel časopisu *Ikaros* vyvíjí nový publikační systém, který bude umožňovat generování metadatových záznamů ve struktuře Dublin Core do zdrojového kódu článků.

Registrované seriály lze rozdělit do tří skupin:

#### 1. seriály vycházející výhradně v elektronické formě

Početně největší skupina, do níž jsou zařazeny zpravodajské servery a další dynamické webové informační systémy, které jsou zpravidla aktualizovány průběžně a které tradiční klasifikace seriálů nezná – např. Česká škola, Živě (viz příloha F.15), root.cz, Svět Namodro, Instantní Astronomické noviny aj. Kromě nich jsou v této skupině zastoupeny seriály, které do značné míry napodobují tradiční vydavatelský

model – mají předem stanovenou periodicitu. Patří k nim např. Ikaros (viz příloha F.12), Chirurgie aj.

## 2. online přílohy tištěných seriálů

Rostoucí skupina seriálů, jejichž charakteristickým znakem je částečná obsahová nezávislost na „originálu“ a které mají z technického hlediska rovněž podobu dynamických databází – např. EkoList po drátě (viz příloha F.14), iDNES, iHNed aj.

## 3. elektronické verze tištěných seriálů

Těmito seriály nemá smysl se z praktického i metodického hlediska zabývat, protože jde o zpřístupnění téhož dokumentu jiným způsobem (viz příloha F.13). Často mají tyto seriály sloužit k marketingové podpoře svých tištěných protějšků, někdy elektronické verze nejsou volně přístupné.

### **B.2.2 Oblast problematiky informačních a komunikačních technologií**

Ve druhém roce řešení byl server projektu WebArchiv vybaven záložním zdrojem (UPS) a probíhalo na něm testování a vývoj stávajících i mnoha nových softwarových nástrojů (dostupné na <http://webarchiv.nkp.cz>):

#### **Dublin Core Metadata Generator**

Tento nástroj (viz příloha F.1) byl původně převzat s minimálními úpravami – ty se soustředily především na jeho zprovoznění v prostředí vývojového serveru a na jeho převezení do češtiny. V testovacím provozu byly záhy odhaleny některé nedostatky.

Například při analýze a extrakci metadat z webových stránek bylo nekorektně nakládáno s opakovanými prvky a nebylo tedy možné využít generátor pro editaci již publikovaných metadat. Pokud byla metadata uložena mimo vlastní HTML soubor ve formátu *xml*, byla pro generátor zcela nedosažitelná.

Drobné chyby byly odhaleny i v chování dynamického formuláře pro tvorbu metadat.

Všechny tyto nedostatky byly odstraněny a zároveň byla výrazně rozšířena funkčnost programu.

Analytická část byla rozšířena o podporu formátů Adobe Acrobat (*pdf*) a Microsoft Word (*doc*). Byly také odděleny funkce analýzy dokumentu a extrakce stávajících metadat Dublin Core. Tak byla zajištěna možnost opakované editace již publikovaných metadat.

V části formuláře byla doplněna funkce poloautomatického přidělení URN (Uniform Resource Name) – na jedno kliknutí myši, která zajišťuje uživateli větší pohodlí a výrazně zmenšuje riziko chyb.

Nabídka výstupních formátů byla aktualizována tak, aby byla vygenerovaná metadata kompatibilní s XHTML 1.0 v případě výstupu do HTML a aby byl zpřesněn a zpřehledněn výstup do XML/RDF (viz příloha F.2).

Vnitřně byl celý program zpřehledněn, byly odstraněny některé jeho nadbytečné části a proměnné a byl synchronizován obsah české a anglické verze. Zároveň byl generovaný HTML kód formuláře upraven tak, aby odpovídal XHTML 1.0.

### **Generátor URN**

Tento jednoduchý nástroj (viz příloha F.3) pro přidělování jednoznačného čísla národní bibliografie pro elektronické dokumenty byl drobně upraven tak, aby mohl přímo spolupracovat s DC generátorem. Jeho kód byl oproti původní finské verzi poněkud pročištěn a program byl plně lokalizován.

### **Kalkulátor MD5**

Tento nástroj (viz příloha F.4) umožnil vyplnit mezeru existující v nabídce služeb serveru WebArchiv. Ačkoli je primárním úkolem tohoto nástroje být pomůckou správce nebo badatele v digitálním archivu, je jeho využitelnost mnohem širší.

Na Internetu existuje několik míst, která umožňují vypočítat kontrolní součet MD5 zadaného řetězce znaků. Kalkulátor provozovaný na serveru WebArchiv jde ale ještě o krok dál: pokud je zadaným řetězcem znaků platné URL, dokáže spočítat nejen kontrolní součet tohoto URL, ale i kontrolní součet dokumentu, který se na daném URL právě nachází. Uživatel by pak mohl tohoto kontrolního součtu využít ke zjištění, jak je daný dokument v archivu zastoupen nebo zda se nějak liší od zdánlivě stejného dokumentu na jiné adrese.

### **Nedlib Harvester**

Tento nástroj (viz příloha F.5) byl ústředním bodem letošních výzkumných a vývojových prací, jelikož právě tento nástroj bude sloužit k automatickému doplňování digitálního archivu. Po několika částečně neúspěšných pokusech se podařilo jej zprovoznit natolik, že mohlo být započato zkušební stahování (*harvesting*). To bylo využito k otestování chování harvesteru v podmínkách reálného provozu; přineslo zajímavá zjištění i odhalilo některé chyby a problémy získané z rutinního provozu harvesteru.

Nejzávažnějšími chybami harvesteru bylo občasné padání nebo zamrzávání některých jeho procesů. Tyto nedostatky byly odstraněny a byl o nich informován autor harvesteru, Mika Rissanen. Tím je zajištěno, aby se tyto chyby v nových verzích harvesteru neopakovaly.

Zbývajícím největším problémem byla taková úprava celého harvesteru, která by umožnila na jednom počítači spustit nezávisle na sobě dvě instance harvesteru. Bylo proto nutné odstranit z programu některé napevno zakompilované parametry, jako pevné názvy cest k souborům, přístup k databázi pevně daného jména apod. Tyto úpravy jsou v současnosti hotové a ačkoli je ještě bude nutné poskytnout autorovi harvesteru do Finska k zapracování do aktuální vývojové verze, lze již nyní hovořit o tom, že bude možné naplnit záměr archivovat vybraná webová periodika (dynamické dokumenty), jejichž obsah je velice proměnlivý, častěji, než zbytek českého webu.

Vývojové práce na harvesteru byly zaměřeny samozřejmě i na další aspekty provozu harvesteru.

Aby byla usnadněna obsluha harvesteru, bylo vyvinuto webové rozhraní pro jeho konfiguraci a ovládání. Toto rozhraní se v současné verzi skládá z několika webových stránek:

**Restrictions** – pomocí tohoto rozhraní je možno nastavit pravidla pro stahování na základě URL – lze zakázat nebo povolit přístup k serverům na základě příslušnosti k určité doméně, určitému serveru, nebo cesty k souboru (viz příloha F.6). Správnost nastavených omezení lze testovat zadáním URL, které je pak proti těmto omezením ověřeno.

**Config** – nastavení dalších okrajových podmínek (viz příloha F.7). Pomocí tohoto rozhraní lze definovat hloubku zanoření stahovaných souborů v rámci serveru, respektování nastavení souboru robots.txt, podporu protokolu ftp a logování zamítnutých URL.

**Statistics** – nástroj pro zjištění jednoduchých statistik probíhající „sklizně“ (viz příloha F.8).

Bohužel, v reálném provozu se ukázalo, že objem a složitost databáze harvesteru neumožňuje klást pomocí webového rozhraní složitější dotazy, protože jejich zodpovězení by trvalo neúnosně dlouho (až desítky minut) a pravidelně končí vypršením času (*timeout*).

Při první, testovací „sklizni“ českého webu, která proběhla v měsících září a říjen 2001, bylo staženo celkem 129 GB dat, a to i přesto, že stahování šlo do hloubky jen 25 zanoření a nebyly brány v potaz soubory na ftp serverech, ani dynamicky generované stránky s parametry. Příložená tabulka, vytvořená přibližně v polovině testu, ukazuje poměrné zastoupení jednotlivých typů souborů. Protože databáze harvesteru neobsahuje údaj o MIME typu jednotlivých souborů, byly soubory rozděleny podle přípon v jejich názvech, což není zcela přesné, ale postačuje pro predikci budoucího složení databáze a poměrného zastoupení jednotlivých typů souborů v ní obsažených a tím i pro odhad diskové kapacity, potřebné pro tvorbu indexů při zpřístupňování archivovaných dokumentů (viz příloha F.9). V každém případě se potvrdila zkušenost severských zemí, že přes 90 procent dokumentů je tvořeno soubory typu *jpg* (fotografie, obrázky), *gif* (grafika webových stránek) a *html* (hypertext). Na druhou stranu je zde již znatelný nárůst počtu souborů *mp3* (hudba, zvuk) a *mpg* (video). Podrobnější statistiky a zkoumání dat v archivu obsažených bude však muset, přes svou zajímavost, počkat až do doby, kdy bude vytvořena podpůrná infrastruktura pro indexování a zpřístupnění archivu, protože prozatím vytvořené nástroje slouží jen k základní orientaci v archivu:

**Tar parser** umožňuje procházet jednotlivými balíky dokumentů vytvořených harvesterem a prohlížet si dokumenty a metadata v nich obsažená (viz příloha F.10).

**URL locator** (viz příloha F.11) naopak nabízí možnost jednoduchého vyhledávání podle URL a podle času stažení dokumentu, přičemž ovšem využívá databáze harvesteru, která není pro tyto účely určena – její obsah se týká vždy jen aktuální „sklizně“.

V současné době nejkritičtějším místem, na které se bude muset soustředit další vývoj, je z hlediska ostrého nasazení modul Packer, který má za úkol stažené soubory a

jejich metadata ukládat v archivačních balících po stech souborech a v komprimované podobě (TAR.GZ). V současné době dochází právě při této operaci k největšímu riziku chyby a následné ztrátě dat. Bude také potřeba zabezpečit bezrizikové pokračování „sklizně“ při jejím přerušení (nyní při spuštění harvesteru bez parametrů dojde k vymazání databáze!).

Koncem roku 2001 byl na Matematicko-fyzikální fakultě UK vypsán ročníkový týmový vývojový projekt na vytvoření vyhledávací infrastruktury pro WebArchiv, do kterého se zapojilo několik studentů MFF UK (projekt vede zástupce spolupracujícího ÚVT MU). Tato infrastruktura by měla zpřístupnit stažené dokumenty v jejich kontextu, tedy s vloženou grafikou ze stejné doby a s odkazy vedoucími primárně opět do archivu na dokumenty ze stejného období. Vyhledávání by mělo být umožněno nejen na základě URL nebo MD5, ale i na základě z dokumentu extrahovaných metadat nebo fulltextového vyhledávání. Tato infrastruktura by měla podporovat (byť nutně omezené) vyhledávání přes Z39.50 a měla by být otevřená tak, aby bylo možné kdykoli připojit další moduly pro indexování netextových typů souborů. (Jakkoli se to může zdát na první pohled nereálné, nástroje tohoto typu již existují. Jeden z nich, *RetrievalWare*, je ve vlastnictví Národní knihovny ČR a jedním z cílů zmíněného ročníkového projektu bude pokus o jeho využití pro indexování archivu).

### B.3 Posun znalostí

K výraznému posunu znalostí došlo zejména v těchto oblastech:

- stanovení kritérií výběru zdrojů pro získávání do digitálního archivu a pro registraci v České národní bibliografii,
- legislativní aspekty - přehled situace v uzákonění povinného výtisku pro elektronické online dokumenty v nejvýznamnějších zemích a související problematiky autorskoprávní; pracovní návrh dohody mezi depozitní knihovnou (NK ČR) a vydavatelem jako prozatímního řešení pro oprávnění nakládat s elektronickými síťovými publikacemi,
- aplikace a další vývoj technických nástrojů pro činnosti související se získáváním, ochranou a zpřístupňováním elektronických zdrojů při dodržení autorskoprávních podmínek,
- vytvoření základních podmínek pro postupné zajišťování problematiky registrace, ochrany a zpřístupňování elektronických zdrojů v provozních podmínkách.

Souhrnně se dá konstatovat, že zatímco v prvním roce řešení projektu byl výzkum na úrovni teoretické, ve druhém roce již přešlo řešení do praktické fáze.

## C NÁVRHOVÁ ČÁST

### C.1 Výsledky řešení

V průběhu řešení dvouletého pilotního projektu (testování) byly vytvořeny základní podmínky, resp. předpoklady pro postupné zajišťování problematiky registrace, ochrany a zpřístupňování elektronických zdrojů v provozních podmínkách.

#### C.1.1 Spolupráce s vydavateli

Již v prvním roce řešení projektu byla navázána spolupráce s vytypovanými vydavateli, další vydavatelé byli kontaktováni v roce 2001. Tito vydavatelé (přehled viz příloha F.19) byli požádáni o testování generátoru metadat Dublin Core (blíže o tomto nástroji viz kapitola B.2.2; ukázka viz příloha F.1), vytváření záznamů v metadatovém schématu Dublin Core pro online elektronické zdroje vydávané těmito institucemi (s využitím zmíněného generátoru) a vložení vytvořených záznamů do zdrojového kódu příslušných dokumentů (viz příloha F.16). Výsledky této části řešení projektu (tj. přehled webových zdrojů s vloženým záznamem metadat Dublin Core) jsou dostupné na internetové adrese <http://webarchiv.nkp.cz/dc.php> (databáze přístupná na této adrese je příležitostně doplňována o další údaje).

V polovině roku 2001 uspořádali řešitelé pro spolupracující vydavatele seminář, jehož smyslem bylo vysvětlit celkově problematiku práce s internetovými zdroji, předvést vyvinuté softwarové nástroje a seznámit s jejich funkcí a používáním v procesu publikování elektronických zdrojů v síti Internet tak, aby tyto nástroje pomáhaly při vyhledávání publikovaných zdrojů. Dále bylo záměrem semináře přiblížit související legislativní problematiku a nastínit návrhy jejího řešení.

#### C.1.2 Legislativa

Zákony zabývající se povinným výtiskem monografií a periodik, které jsou v současné době v platnosti v České republice (zákon č. 37/1995 Sb. „o neperiodických publikacích“ a zákon č. 46/2000 Sb. „tiskový zákon“), nelze ve stávající podobě aplikovat na elektronické zdroje přístupné online. Bude proto třeba provést právní rozbor těchto zákonů a následně zpracovat návrhy na jejich změnu.

Pro přechodné období by měla vejít v platnost dohoda s vydavateli týkající se dobrovolného odevzdávání výtisků online elektronických dokumentů do konzervačního fondu Národní knihovny ČR a následného zpřístupňování těchto zdrojů uživatelům Národní knihovny. Řešitelé projektu vypracovali zatím pracovní návrh takové dohody. Hlavním podkladem pro tuto dohodu byl dokument *Mezinárodní deklarace k odevzdávání elektronických dokumentů do konzervačního fondu (International declaration on the deposit of electronic publications)*, kterou společně vytvořily *Conference of European*

*National Librarians a Federation of European Publishers* (český překlad byl pořízen v rámci řešení projektu a je součástí příloh ke zprávě – viz příloha F.20; připravuje se též jeho publikování v časopise Národní knihovna), a podobné místní dohody, které jsou již v platnosti v Nizozemí, ve Velké Británii a v Německu.

Na základě zkušeností s praktickým prováděním této dohody, které by měly být pravidelně vyhodnocovány, budou vypracovány jednak případné změny a doplnění dohody, jednak již zmíněné návrhy na změnu legislativy týkající se povinných výtisků.

### **C.1.3 Informační a komunikační technologie**

Vytvořená infrastruktura již v současné podobě vyhovuje pro archivaci českého webu, ovšem její vývoj (v podstatě stejně jako v případě všech softwarových produktů) nemůže být nikdy zcela ukončen. Zde nejde jen o hledisko potřeb uživatele nebo provozovatele, ale i o hledisko technického vývoje nebo o legislativní problematiku.

Vývoj Generátoru metadat bude muset odrážet změny v nárocích uživatelů a změny technického rázu (podpora dalších vstupních a výstupních formátů, proměny kvalifikátorů a možná i samotného Dublin Core).

Generátor URN bude muset být přizpůsoben pro automatizované přidělování URN jiným softwarovým produktům nejen na stejném serveru, ale i po síti.

Kalkulátor MD5 bude pravděpodobně integrován přímo jako jeden ze vstupních bodů do archivu.

Harvester bude nadále vyvíjen ve Finsku při spoluúčasti národních knihoven dalších zemí včetně Národní knihovny ČR.

#### **Návrhy pro další vývoj v oblasti ICT:**

Stávající hardwarová platforma je pro ostrý provoz nevyhovující.

Bude nutno optimalizovat rychlost komunikace serveru WebArchiv a serveru *lenoch*, ke kterému je připojen páskový robot (technika pořízená v NK v rámci projektů zaměřených na digitalizaci), na němž jsou uložena veškerá archivovaná data. Dále bude třeba zajistit solidnější hardwarovou platformu, než je současný PC server.

Pro reálný provoz nástroje *RetrievalWare* je stávající hardware nevyhovující. To je dáno jednak nemožností souběhu harvestingu a indexace na jednom serveru (hrozící přetížení systémových prostředků) a jednak tím, že filesystem linuxového serveru nepodporuje soubory větší než 2GB. Přitom jen fulltextové indexy HTML souborů budou pravděpodobně přesahovat už nyní tento limit až pětkrát. Také server, na kterém je provozován *RetrievalWare*, bude muset být posílen, protože jeho paměť je nyní jen 512MB. Dále bude potřeba rozšířit jeho diskové pole tak, aby bylo možné vytvořit potřebné indexy.

## C.2 Propagace projektu

Veškeré informace týkající se projektu řešitelé průběžně publikují na webových stránkách projektu na adrese <http://webarchiv.nkp.cz>. Vedle české verze stránek je vypracována také verze anglická.

V jednotlivých sekcích se nacházejí

- dokumenty, které vznikly v rámci řešení projektu nebo souvisejí s jeho předmětem, a jejich autory jsou řešitelé projektu
- odkazy na relevantní zdroje (zejména elektronické) a zahraniční projekty obdobného zaměření
- generátor metadat Dublin Core, generátor URN a kalkulátor MD5

Odkazy na webové stránky projektu jsou umístěny v nejvýznamnějších českých vyhledávacích službách (katalozích) Internetu.

Informace o projektu, včetně odkazu na webové stránky, jsou zveřejněny na webových stránkách *Dublin Core Metadata Initiative* (<http://www.dublincore.org>), v sekci „Dublin Core Projects“.

Řešitelé představili problematiku řešenou v rámci projektu i projekt samotný na řadě seminářů a konferencí (většina příspěvků přednesených na těchto akcích je umístěna na webových stránkách projektu, v sekci „Dokumenty“):

<i>datum</i>	<i>název akce</i>	<i>místo konání akce</i>
5/2000	Inforum 2000	Vysoká škola ekonomická v Praze
11/2000	RUFIS 2000	Vysoké učení technické v Brně
4/2001	Automatizace knihovnických procesů 2001	SVK v Liberci
6/2001	seminář Dublin Core 2001	Národní knihovna ČR
9/2001	Knihovny současnosti 2001	Seč u Chrudimi
9/2001	rekvalifikační kurz	Národní knihovna ČR
9/2001	Automatizace knihoven 2001	areál Univerzity Karlovy, Praha – Jinonice
11/2001	Moderní inf. a komunikační technologie v knihovnictví	Státní technická knihovna

Dále byla veřejnost seznámena s problematikou řešenou v rámci projektu i s projektem samotným v řadě článků (viz rešerše v úvodu této zprávy).

V neposlední řadě řešitelé projektu připravili leták, který stručně informuje o projektu, metadatech a metadatovém standardu Dublin Core (viz příloha F.17, F.18). Leták byl distribuován na seminářích a konferencích, kterých se řešitelé projektu v průběhu roku zúčastnili.

### C.3 Závěr

Cíle řešení projektu byly splněny. V podmínkách testování byly vytvořeny předpoklady pro postupné zavádění provozního zpracování této agendy. Konkrétně byly připraveny podklady pro právní zabezpečení získávání, archivace a zpřístupňování domácích elektronických zdrojů publikovaných v síti Internet, softwarové nástroje pro provádění těchto činností a byla navázána spolupráce s vybranými vydavateli síťových elektronických zdrojů pro simulaci těchto činností v praxi. Data získaná v rámci automatizovaného stahování zdrojů z webu mohou být využívána rovněž pro registraci elektronických zdrojů dostupných online v České národní bibliografii.

Od vytvoření základních předpokladů v podmínkách testování k provoznímu řešení problematiky trvalého zajištění ochrany a zpřístupnění síťových elektronických zdrojů je však ještě dlouhá a náročná cesta, vyžadující značné finanční prostředky zejména na investiční vybavení (hardware) a jeho průběžné obnovování i na průběžnou aktualizaci softwarových nástrojů. Rovněž je třeba počítat s nároky na lidskou práci související jak s tvorbou bibliografické databáze, tak s řízením výpočetní a komunikační techniky.

Pilotní projekt byl plánován na dva roky (2000 - 2001) a vzhledem k tomu, že byl zrušen roční program VaV na rok 2002 *Zpřístupňování a ochrana knihovních fondů formou digitalizace s využitím mezinárodní sítě Internet v souvislosti s vytvářením informační společnosti*, do něhož podávali řešitelé v roce 2001 žádost o grant, nemá tento projekt prozatím pokračování. Národní knihovna ČR může se stávající technikou zajistit pokračování činností alespoň na minimální úrovni, aby dosažené výsledky projektu zcela nezapadly. Pro překlenovací období by mohl pomoci grant z programu VISK3, do něhož podali řešitelé návrh projektu počátkem ledna 2002. Pokud bude tento projekt v rámci programu VISK3 schválen, měl by být součástí jeho řešení odhad věcných a finančních předpokladů (hardware a software) pro průběžnou tvorbu a zpřístupňování webového archivu a pro „údržbu“ archivovaných zdrojů (technologie migrace dat, emulace aj. - v souvislosti s morálním stárnutím nástrojů interpretace elektronických zdrojů). Problematiku digitálních zdrojů je třeba řešit v kontextu s problematikou digitalizovaných dat, s níž má zejména po technické stránce mnoho společného, tj. jako komplex digitální knihovny.

### C.4 Návrhy opatření

1. Najít zdroje financování pro pokračování projektu v roce 2002 a dalších letech s cílem zajištění trvalé ochrany a realizace bibliografické kontroly v oblasti elektronických dokumentů dostupných online.
2. Společně s digitalizovanými dokumenty a dalšími zdroji informací v digitální podobě vytvářet postupně digitální knihovnu a umožnit zpřístupňování dokumentů, resp. dat z digitální knihovny prostřednictvím jednotné informační brány.

## D RESUMÉ A KLÍČOVÁ SLOVA

### D.1 Resumé

Problematika shromažďování, dlouhodobé ochrany a zpřístupňování síťových elektronických zdrojů zahrnuje otázky knihovnické a legislativní, které jsou podmíněny řešením informačních a komunikačních technologií. Ve světě se této komplexní problematice věnuje pozornost na národní i mezinárodní úrovni již od počátku 90. let minulého století. V České republice i v zemích bývalého východního bloku byla v době zahájení projektu tato problematika zcela nová.

Cílem projektu bylo připravit podmínky pro zpracování české národní bibliografie elektronických zdrojů, se zaměřením zejména na zdroje dálkově přístupné. S bibliografickým zpracováním souvisí zajištění trvalého uchování domácích elektronických dokumentů monografických i seriálových, publikovaných v síti Internet, a jejich zpřístupnění, respektující autorská práva.

V rámci projektu byla provedena analýza řešení této komplexní problematiky ve světě a na jejím základě se řešitelé orientovali zejména na řešení v evropských severovýchodních zemích, kde jsou výsledky jak v oblasti knihovnické a legislativní, tak i v oblasti technické velmi progresivní a zároveň jsou kulturně-sociální podmínky i technologické podmínky srovnatelné se stavem u nás. Pro průběžné testování prací a softwarových nástrojů byl pořízen unixový server, který sloužil k instalování nástrojů pro stahování dokumentů, pro ukládání údajů pro popis zdrojů aj. a pro ukládání zdrojů do webového archivu. Na tomto serveru byla také zřízena webová prezentace projektu (na adrese <http://webarchiv.nkp.cz>).

Výsledky řešení lze shrnout do následujících bodů:

- Byla stanovena kritéria výběru zdrojů pro získávání do digitálního archivu a pro registraci v České národní bibliografii.
- Byla zmapována situace v uzákonění povinného výtisku pro elektronické online dokumenty a související problematika autorskoprávní a byl zpracován pracovní návrh dohody mezi depozitní knihovnou (NK ČR) a vydavateli jako prozatímního řešení pro oprávnění nakládat s elektronickými síťovými publikacemi.
- Byly aplikovány zahraniční softwarové nástroje, lokalizovány pro domácí podmínky a byl prováděn jejich další vývoj.
- Byly vytvořeny základní podmínky pro postupné zajišťování problematiky registrace, ochrany a zpřístupňování elektronických zdrojů v provozních podmínkách.

### D.2 Klíčová slova

*národní bibliografie \* ochrana knihovních fondů \* elektronické zdroje \* Internet \* elektronický archiv \* legislativní dokumenty \* povinný výtisk \* autorské právo \* akvizice \* zpřístupňování dokumentů \* indexace \* mezinárodní standardy \* metadata \* Dublin Core*

## E PŘÍLOHY

- E.1 [Generátor metadat Dublin Core \(lokalizovaná verze\)](#)
- E.2 [Výstup z generátoru metadat Dublin Core ve formátu XML/RDF](#)
- E.3 [Generátor jednoznačného identifikátoru URN](#)
- E.4 [Kalkulátor kontrolního součtu MD5](#)
- E.5 [Funkční schéma nejnovější verze programu Nedlib Harvester](#)
- E.6 [Základní pravidla pro stahování na základě URL pomocí programu Nedlib Harvester](#)
- E.7 [Díličí pravidla pro stahování pomocí programu Nedlib Harvester](#)
- E.8 [Statistiky stahování pomocí programu Nedlib Harvester](#)
- E.9 [Zastoupení formátů webových zdrojů v archivu](#)
- E.10 [Obsah jednoho archivního balíku souborů získaných pomocí Nedlib Harvesteru](#)
- E.11 [Vyhledávací rozhraní v archivu webových zdrojů](#)
- E.12 [Elektronický seriál s pevnou periodicitou \(Ikaros\)](#)
- E.13 [Elektronická verze tištěného seriálu \(Forum\)](#)
- E.14 [Elektronická příloha tištěného seriálu \(EkoList po drátě\)](#)
- E.15 [Elektronický seriál průběžně aktualizovaný \(Živě.cz\)](#)
- E.16 [Článek z elektronického seriálu s metadatovým záznamem podle Dublin Core \(Česká škola\)](#)
- E.17 [Leták se základními informacemi o projektu WebArchiv a Dublin Core \(přední strana\)](#)
- E.18 [Leták se základními informacemi o projektu WebArchiv a Dublin Core \(zadní strana\)](#)
- E.19 [Přehled vydavatelů spolupracujících na tvorbě záznamů Dublin Core](#)
- E.20 [Překlad textu „International declaration on the deposit of electronic publications“](#)