DLOUHODOBÁ OCHRANA DAT V NÁRODNÍ LÉKAŘSKÉ KNIHOVNĚ – AKTUÁLNÍ STAV A PROBLÉMY

Lenka Maixnerová, Filip Kříž, Ondřej Horsák, Helena Bouzková

Anotace

Národní lékařská knihovna (NLK) provozuje od roku 2008 digitální archiv v systému Kramerius. Archiv aktuálně obsahuje již přes 290 digitálních dokumentů, které byly získány z různých zdrojů, různými způsoby a na základě různých licencí. V roce 2009 NLK zahájila spolupráci s autory a vydavateli vědeckých publikací, které po uzavření licenční smlouvy mohou být archivovány a zpřístupněny v digitálním archivu. Byly získány cenné zkušenosti při provozování archivu, zejména v oblasti získávání dat od vydavatelů, která se ukazuje jako klíčová pro efektivní zpracování a zpřístupnění. V rámci pokračujícího vývoje portálu Medvik bylo mimo jiné zprovozněno dynamické propojení databáze Bibliographia medica Čechoslovaca s archivovanými časopisy a sborníky, takže se uživatel snadno dostane k plnému textu. NLK dosáhla v oblasti digitalizace hmatatelných výsledků, ale je třeba se systematicky více zaměřit na dosažení parametrů důvěryhodného digitálního archivu, které jsou definovány standardem OAIS.

Klíčová slova

informační služby, digitální archivy, Národní lékařská knihovna, uchovávání dokumentů, digitalizace, zdravotnické knihovny, elektronické dokumenty, digital preservation

Úvod

Získávání, zpracování, trvalé uchování a zpřístupnění fondů a sbírek tradičních dokumentů je v prostředí knihoven relativně dobře a kvalitně zajištěno. Knihovny se však musí vypořádat s hrozbou ztráty tištěných dokumentů způsobenou degradací kyselého papíru, který se používal téměř 150 let a poškozením dokumentů častým používáním. Efektivním prostředkem pro uchování ohrožených publikací je jejich digitalizace a následné zpřístupnění v digitálních knihovnách, archivech nebo repositářích. V souvislosti s rozvojem informačních a komunikačních technologií knihovny řeší získávání, zpracování, uložení a zpřístupnění nových typů dokumentů, které vznikly již elektronicky. Zejména se jedná o problémy dlouhodobého uchování a zpřístupnění těchto digitálních dokumentů.

Potřeba a nutnost dlouhodobě uchovávat a zpřístupňovat kulturní dědictví a vědecké informace v tradičních i elektronických dokumentech je formulována v iniciativě Komise evropských společenství i2010: Digital Libraries Initiative [1]. Pro knihovny v České republice je určena Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010, která předpokládá vytváření oborových digitálních knihoven a archivů s návazností na systém Národní digitální knihovny [2].

Situace v NLK

Národní lékařská knihovna (NLK) se již v roce 2004 zapojila do programu MK ČR VISK 6 „Národní program digitálního zpřístupnění vzácných dokumentů Memoriae Mundi Series Bohemica" a v roce 2008 do programu MK ČR VISK 7 „Národní program mikrofilmování a digitálního zpřístupňování dokumentů ohrožených degradací kyselého papíru – Kramerius". Následným logickým krokem je zpřístupnění výstupů uvedených programů uživatelům NLK. Knihovníci a informační pracovníci NLK rozvíjejí aktivity v získávání dokumentů a jejich digitalizace dokumentů z fondu NLK. Budování a údržba digitálního archivu (DA) je perspektivním úkolem NLK a vyžaduje investiční finanční prostředky. NLK zahájila spolupráci s českými nakladateli odborných biomedicínských periodik, která se soustřeďuje na možnosti získávání a zpřístupnění elektronických dokumentů v digitálním archivu. Neméně důležité je získávat obsah archivu přímo od autorů vědeckých informací a motivovat je k využívání DA pro archivaci a zpřístupňování jejich prací.

Digitální archiv NLK

Pro implementaci DA byl vybrán český systém Kramerius [3]. Kramerius je software s otevřeným zdrojovým kódem produkovaný firmou Qbizm Technologies, na jehož vývoji se podílí Národní knihovna ČR (NK ČR), Knihovna Akademie věd ČR a Moravská zemská knihovna v Brně. Vývoj systému je zajišťován finančními dotacemi Ministerstva kultury ČR a Ministerstva školství, mládeže a tělovýchovy ČR. Systém Kramerius slouží ke zpřístupňování digitálních dokumentů v souladu s autorským zákonem. V současnosti podporuje periodika a monografie, do budoucna se plánuje rozšíření i pro ostatní typy dokumentů. Systém umožňuje replikaci dat mezi jednotlivými instalacemi.

Provoz DA zahrnuje tyto činnosti:

1. Akvizice - výb ěr a získání dokumentů pro DA
2. Zpracování
3. Ochrana
4. Zpřístupnění

Akvizice

Digitalizace tištěných periodik a monografií se odvíjí zejména od možností finančního rozpočtu NLK, neboť v programech VISK je vždy vyžadována spoluúčast příjemce. U vlastní digitalizace jsme omezeni kapacitou zpracovatelské linky a počtem digitalizačních pracovišť - zatím pouze 1. Digitalizují se primárně poničené exempláře vytištěné na kyselém papíru a často využívané příručkové publikace, které nemá NLK ve více výtiscích. Před zahájením digitalizace se vždy ověřuje, zdali daný dokument nedigitalizovala již jiná instituce v ČR. Pro tento účel je možné využít Registr digitalizace [4]. U dokumentů, které vznikly již elektronickou cestou je klíčové zahájit spolupráci

s nakladateli a autory. Toto se ukazuje jako problém - řada nakladatelů a autorů má neopodstatněnou obavu z podepsání smlouvy, která je podmínkou pro trvalou archivaci a zpřístupnění dokumentů v DA. Dalším zdrojem může být replikace digitalizovaných dokumentů, které má NLK ve fondu, z archivů jiných knihoven ČR na základě vzájemné dohody, zejména s Národní knihovnou ČR.

Vlastní digitalizace
Dokument určený k digitalizaci je komplet naskenován v rozlišení 300 DPI (Plustek OpticBook 4600) a je automaticky provedeno OCR, výstupem je soubor ve formátu PDF s textovou vrstvou. Po kontrole kvality se následně digitální kopie rozdělí na jednotlivé soubory - části podle typu dokumentu a připojí se k bibliografickému záznamu v systému Medvik. Po doplnění a kompletaci technických a administrativních metadat je vyexportován XML soubor dle DTD Kramerius, který je následně importován do DA.

Zpracování
Pro všechny digitalizované dokumenty je vytvořen (pokud již neexistuje) bibliografický záznam v systému Medvik [5]. Digitalizovaný dokument v DA a bibliografický záznam jsou vzájemně propojené přes URL odkazy, takže se uživatel dostane z bibliografického záznamu přímo do digitální verze a naopak. V případech velmi zničených exemplářů, mají uživatelé k dispozici už pouze digitální kopii. U dokumentů, u nichž se provádí analytické zpracování (periodika, sborníky, grantové zprávy) se propojují jednotlivé části digitálního dokumentu s databází Bibliographia medica Čechoslovaca (BMČ). Protože nejsou metadata v BMČ a v archivu vždy strukturálně shodná, byl vytvořen dynamický linkovací mechanismus, který uživatele odkáže ze záznamu článku v BMČ na nejbližší existující úroveň metadat v digitálním archivu (úrovně: článek - číslo - ročník - titul).

Zpracování dat z VISK a Manuscriptoria
Externě digitalizované dokumenty (program VISK 7) jsou získávány již ve formátu Kramerius - po kontrole a případných opravách metadat jsou data importována do DA. Pro potřeby konverze metadat ve formátu Manuscriptorium (VISK 6) byl vytvořen konverzní program, po nezbytných úpravách metadat a konverzi obrazových souborů na formát DjVu mohou být data importována do DA.

Zpracování dat od vydavatelů
Zpracování dat od vydavatelů velmi závisí na možnostech konkrétního vydavatele poskytnout digitální dokumenty, případně i metadata. Obvykle jsou získány soubory ve formátu PDF, které obsahují jednotlivé články nebo čísla časopisu, ke kterým je třeba metadata vygenerovat z katalogu Medvik a BMČ. Metadata jsou obvykle předávána jako XML soubory v nestandardních formátech. V některých případech se nepodařilo od vydavatele získat žádná data, takže jsme byli odkázáni pouze na webovskou stránku časopisu s plnými texty. V takovém případě byl použit software WebHarvest [6], který umožňuje

stáhnout potřebná data přímo ze stránek vydavatele podle nastaveného profilu. Takto získaná data (metadata, PDF soubory, obrazové soubory, HTML stránky) jsou následně zpracována na importní balíček pro digitální archiv. Pokud získaná metadata obsahují alespoň názvy článků, jsou porovnávána se záznamy BMČ a v případě shody jsou obohacena o identifikátor článku BMČ. V případech, kdy stránky vydavatele poskytují pouze HTML obsah, jsou jednotlivé články nebo čísla časopisu kompletována s obrazovými soubory do formátu PDF. V průběhu zpracování se zpravidla objevují chyby v HTML (chybné kódování znakové stránky, neplatné odkazy, chybějící obrázky, nevalidní HTML atp.). Pokud je to možné, tak jsou chyby odstraňovány na straně vydavatele, případně jsou stažená data upravována až v procesu zpracování dat v NLK.

Ochrana

Zajištění dlouhodobé ochrany digitálních dat je komplexní proces, který zahrnuje nejen technické prostředky ochrany, ale také zajištění organizace, financování, personální zabezpečení a řízení všech procesů probíhajících v systému digitálního archivu [7]. Této oblasti se věnuje řada výzkumných projektů - např. PLANETS, CASPAR a DPE. V rámci těchto projektů vznikají praktické nástroje nejen pro provoz ale zejména pro plánování a řízení systémů pro dlouhodobou ochranu digitálních objektů - např.: PLATTER, DRAMBORA, TRAC. NLK plánuje v blízké době tyto nástroje vyzkoušet v praxi a kontextu budovaného DA. Prozatím jsme se soustředili na organizaci základních funkcí archivu a technické zabezpečení provozu a bezpečnosti dat DA, které využívá infrastrukturu datového centra NLK vybudovaného v rámci projektu MEDVIK.

Zpřístupnění

Podmínky zpřístupnění digitalizovaných dokumentů jsou dány ochrannými lhůtami definovanými v Autorském zákoně. Volně přístupná mohou být periodika vydaná po roce 1888. V případě monografií po uplynutí 70 let od smrti autora. V opačných případech mohou být díla zpřístupněna pouze v rámci počítačové sítě NLK. V případě vědeckých a odborných dokumentů však není tento model z hlediska využitelnosti vhodný. NLK proto připravila Licenční smlouvu o užití díla pro autory, kteří mají zájem o archivaci a zpřístupnění svých děl s odborným obsahem, která stanovuje mimo jiné podmínky, za kterých lze dílo zpřístupnit v DA. S vydavateli kteří mají zájem o archivaci publikovaných titulů NLK uzavírá Smlouvu o poskytování elektronických online zdrojů. Licenční smlouvy jsou dostupné z [8]. Začínáme také s archivací dokumentů zveřejněných pod licencemi Creative Commons (CC).

CC jsou licence, které umožňují legální využívání a sdílení autorských děl. CC vychází z toho, že existují lidé, kteří nechtějí využívat všechna práva k duševnímu vlastnictví, která jim zaručuje zákon, ale která mohou omezovat sdílení a využívání autorských děl. CC nabízí různá licenční schémata a držitelé autorských práv si mohou vybrat, jaká z autorských práv k dílu si chtějí

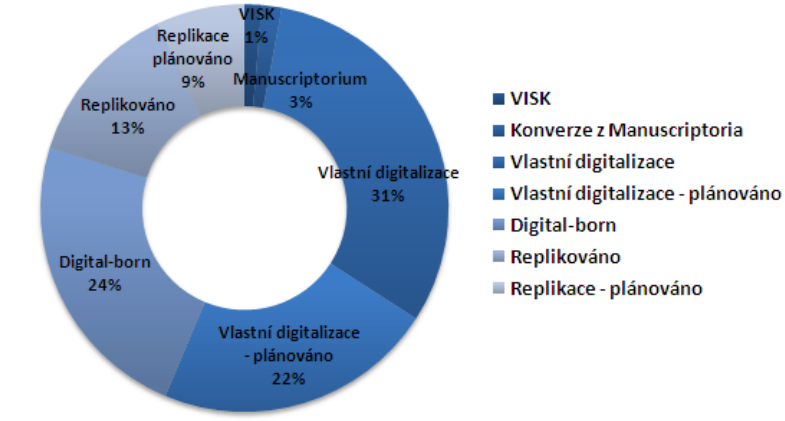
ponechat a jakých se naopak vzdát ve prospěch veřejnosti. Používání CC v ČR je umožněno díky novele Autorského zákona č. 216/2006 Sb. Česká verze CC byla schválena na jaře 2009 [9].

Obsah archivu

Digitální archiv NLK je přístupný na adrese http://www.medvik.cz/kramerius a jeho obsahem jsou a mohou být následující materiály:

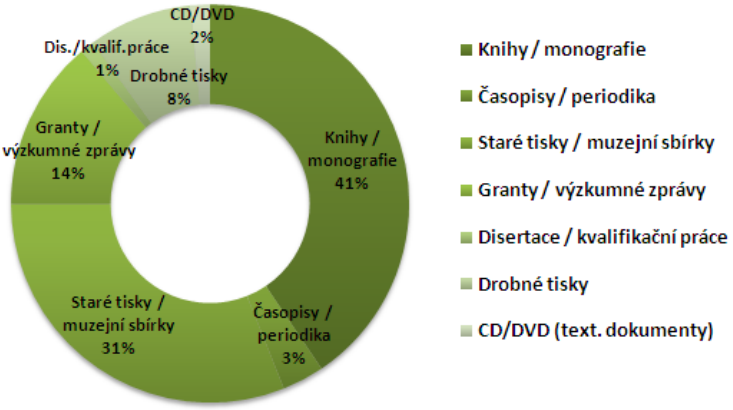
- digitalizovaná tištěná periodika a monografie
- digitalizované staré tisky
- elektronické časopisy
- závěrečné grantové zprávy (IGA MZ ČR)
- konferenční materiály
- digital-born dokumenty

K 1.2. 2010 obsahuje archiv 17 titulů periodik a přes 270 monografických titulů. Nově zpracované dokumenty jsou doplňovány obvykle týdně nebo jednou za 14 dní. Po zpracování plánovaných dokumentů se bude počet dokumentů v DA blížit 400. Aktuální seznam archivovaných dokumentů včetně popisu licencí a podmínek přístupu je uveden v [8 - seznam dokumentů]. Dokumenty byly získány z následujících zdrojů.



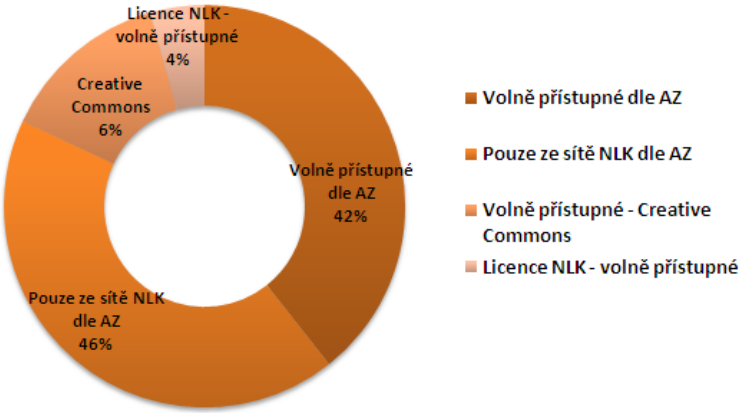
Graf 1 - Obsah archivu podle zdroje dokumentu

Přestože Kramerius standardně podporuje archivaci pouze periodik a monografií, lze ho použít i pro jiné typy strukturou podobných dokumentů. Rozložení obsahu archivu podle fondů NLK je uvedeno v následujícím grafu.



Graf 2 - Obsah archivu podle fondu

Aktuální obsah archivu podle typu licencí je uveden v následujícím grafu.



Graf 3 - Obsah archivu podle typu licence

Digitalizovaná tištěná periodika

Digitalizace tištěných periodik byla v NLK zahájena v roce 2008 díky finanční dotaci z programu MK ČR VISK 7 Kramerius. Pro digitalizaci byla vybrána periodika, tištěná na kyselém papíru. Snahou bylo digitalizovat kompletní vydání (tj. od prvního do posledního vydaného čísla), což se bohužel ne vždy podařilo. NLK též zahájila digitalizaci periodik, které nejsou zničené, ale jsou

často využívané. Pro digitalizaci periodik v rámci programu VISK 7 byla vybrána firma Ampaco ČR, ostatní periodika byla digitalizována přímo v NLK.

Elektronické časopisy
V současné době má většina odborných tištěných časopisů i svou elektronickou verzi, některé již vychází pouze elektronicky. Trvalé a dlouhodobé uchování a zpřístupnění těchto elektronických verzí se stává stále důležitějším úkolem. NLK počátkem roku 2009 oslovila české nakladatele odborných biomedicínských časopisů, zda by měli zájem o službu dlouhodobé archivace a zpřístupnění elektronických verzí periodik v digitálním archivu NLK včetně následného propojení s bibliografickou databází Bibliographia medica Čechoslovaca (BMČ).

Staré tisky ze sbírky Zdravotnického muzea
Zdravotnické muzeum je součástí NLK. Spravuje následující sbírky: akologický kabinet, novější nástroje a přístroje, mince a medaile, Würtzova sbírka sošek, další hmotné památky vztahující se k dějinám lékařství, staré tisky, archiv. V DA jsou zatím k dispozici pouze monografie (cca 135 titulů) z různých zdrojů.

Závěrečné grantové zprávy IGA MZ ČR
NLK archivuje závěrečné zprávy Interní grantové agentury Ministerstva zdravotnictví ČR. Zprávy je možné využívat pouze absenčně v prostorách NLK. Pro velký zájem uživatelů začaly být zprávy v roce 2008 digitalizovány. Plné texty digitalizovaných závěrečných zpráv jsou standardně přístupné v souladu s Autorským zákonem pouze z počítačové sítě NLK. V případě souhlasu hlavních řešitelů (podepsáním Licenční smlouvy o užití díla) jsou díla volně přístupná přes Internet. Publikační činnost, která je přílohou závěrečné zprávy, je přístupná pouze z počítačové sítě NLK a je prolinkována do databáze BMČ (Bibliographia medica Čechoslovaca). Digitalizace zpráv probíhá postupně, přednostně jsou digitalizovány práce, u nichž je podepsána licenční smlouva.

Konferenční materiály
Sborníky, prezentace, postery z konferencí, kongresů, workshopů apod. se řadí mezi tzv. šedou literaturu, kterou nelze získat v běžné distribuční síti. V posledních letech řada z nich vychází pouze v elektronické podobě. V současné době zpracováno 15 digital-born sborníků a další čekají na zpracování. Naše zkušenosti ukazují, že zejména starší online sborníky je často obtížné dohledat a pokud máme uložen z dřívějška v katalogu URL odkaz, tak je velmi často již nefunkční.

Digital-born monografie
Stejně jako u periodik vycházejí některá monografická díla i elektronicky nebo jsou pouze v elektronické podobě. Pro zařazení online monografie do DA je nutné podepsat licenční smlouvu s NLK, případně musí být elektronická verze označená zveřejněna pod licencí CC. Dokumenty, u kterých nebylo možno

uzavřít licenční smlouvu a nejsou zveřejněny pod licencí CC, jsou přístupné v souladu s autorským zákonem pouze z počítačové sítě NLK.

Závěr

Další rozvoj DA NLK by měl směřovat k implementaci referenčního modelu OAIS (Open Archival Information System - standard ISO 14721) [9]. Stávající systém DA standardu OAIS vyhovuje pouze v dílčích aspektech. Sledujeme však vývoj nové verze Krameria, která by měla požadavky OAIS splňovat.

Klíčové pro další rozvoj digitálního archivu NLK je směřovat postupně k dosažení i dalších stanovených atributů důvěryhodného digitálního repozitáře [11], který by umožnil následnou certifikaci. Datová kapacita stávajícího úložiště je omezená a se stoupajícím počtem dokumentů ji bude potřeba rozšiřovat. Budování digitální knihovny NLK přináší nové technologické postupy. Zvětšuje se objem digitalizačních prací, např. závěrečných grantových zpráv IGA MZ ČR a dokumentů ohrožených degradací kyselého papíru, u kterých již hrozí jejich nenávratná ztráta.

Důležitým aspektem je také persistentní identifikace objektů v DA pomocí systému identifikátorů (PID), předpokládáme v budoucnu zapojení do národního systému URN, který je řešen pracovní skupinou PID [12] v souvislosti s projektem NK ČR Národní digitální knihovna. Podle posledních informací by měly být první identifikátory založené na číslech České národní bibliografie generovány již v letošním roce. Důležité pro využití v oblasti lékařských informací však je, zda PID navrhne takové schéma, které bude rozšiřitelné o další systémy identifikátorů, v našem případě číslo BMČ.

Aktuálně se však můžeme soustředit na rozvoj spolupráce s vydavateli a autory odborných lékařských publikací, kteří mají zájem o archivaci a zpřístupnění svých publikací a děl v DA NLK. Plánujeme návrh společného standardu pro předávání metadat časopisů, aby bylo možno zefektivnit zpracování dat od vydavatelů. Další důležitou oblastí je průzkum spolupráce se zdravotnickými knihovnami ČR. Digitální knihovna NLK přináší novou hodnotnou službu zdravotnickým uživatelům.

Reference

[1.] STOKLASOVÁ, B. Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010 [online]. [cit. 2010-01-30]. Dostupný z WWW: <http://www.ndk.cz/narodni-dk/publikace/koncepcekniharch.pdf>.

[2.] STOKLASOVÁ, B. Projekt „Národní digitální knihovna" v širším kontextu. [cit. 2010-01-30]. Dostupný z WWW: <http://www.ndk.cz>.

[3.] Qbism. Uživatelský portál systému Kramerius [online]. 2003-2009 [cit. 2010-01-30]. Dostupný z WWW: <http://kramerius.qbizm.cz>.

[4.] Registr digitalizace : evidence digitalizovaných dokumentů a sledování procesu zpracování [online]. [2009] [cit. 2010-01-30]. Dostupný z WWW: <http://sluzby.incad.cz/esp/rdcz>.

[5.] Portál Medvik - katalog Národní lékařské knihovny [online]. [cit. 2010-01-30]. Dostupný

z WWW: <http://www.medvik.cz/medvik/?library=ABA008>.

[6.] WebHarvest [online]. [cit. 2010-01-30]. Dostupný z WWW: <http://web-harvest.sourceforge.net/>.

[7.] ROSENTHAL, Colin, BLEKINGE-RASMUSSEN, Asger, HUTAŘ, Jan. Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER). 1. vyd. Praha : Národní knihovna České republiky, 2009. 51 s. ISBN 978-80-7050-569-4.

[8.] Národní lékařská knihovna. Digitální archiv [online]. Praha : NLK, 2009 , 20.11.2009 [cit. 2010-01-30]. Dostupný z WWW: <http://www.nlk.cz/informace-o-nlk/odborne-cinnosti/digitalni-archiv>.

[9.] Creative Commons Česká republika - České CC licence [online]. 2009 [cit. 2010-01-30]. Dostupný z WWW: <http://www.creativecommons.cz/zakladni-informace-o-cc/ceske-cc-licence>.

[10.] Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS) [online]. Washington, DC : [s.n.], 2002. 148 s. [cit. 2010-01-30]. Dostupný z WWW: <http://public.ccsds.org/publications/archive/650x0b1.pdf>

[11.] RLG/OCLC Working Group on Digital Archive Attributes. Trusted Digital Repositories: Attributes and Responsibilities : An RLG-OCLC Report [online]. Mountain View, CA : RLG, 2002. 70 s. [cit. 2010-01-30]. Dostupný z WWW: <http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf>.

[12.] Pracovní skupina pro perzistentní identifikátory. Specifikace požadavků – Pracovní skupina PID [online]. 2007-2010 [cit. 2010-01-30]. Dostupný z WWW: <http://pid.ndk.cz/specifikace-pozadavku>.

Kontaktní adresa

Mgr. Filip Kříž
Národní lékařská knihovna
Sokolská 54
121 32 Praha 2
Tel: 296 335 940
e-mail: *kriz@nlk.cz*
http://www.nlk.cz