

Digitalizace

úvod do problematiky

Martin Krčál

Otázka na úvod

- Jak získáme elektronický dokument?



Digitalizace

- převod info z analogové do **digitální** (elektronické) podoby
- formy informace
 - ❖ textové
 - ❖ obrazové
 - ❖ zvukové
 - ❖ jejich kombinacích

Proč digitalizujeme

- dostupnost informací
- úspora místa
- ochrana a archivace
- vyhledávání

Základní pojmy

Rozlišení

- rozdělení obrazu na síť pixelů
- **pixel** = jedna barva
- větší hustota (rozlišení) = větší kvalita = větší velikost souboru
- základní jednotka **dpi** (dots per inch)
- tisková kvalita od 300 dpi
- na web od 75 dpi

Barevná hloubka

- počet barev použitých při skenu
- černobílé skeny: čb nebo stupně šedi
- barevné skeny: 24-bitů+
- zdravé oko vnímá okolo 4 mld. odstínů barev

Barevná hloubka - počty barev

Druh obrazu	Počet bitů	Počet barev
černobílý (bitonální, monochromatické)	1	$2^1 = 2$
stupně šedi (grayscale)	8	$2^8 = 256$ odstínů šedi
8-bitový barevný (color)	8	$2^8 = 256$
16-bitový (high color)	16	$2^{16} = 65\,536$
24-bitový (true color)	24	$2^{24} = 16\,777\,216$
32-bitový (super true color)	32	$2^{32} = 4\,294\,967\,296$
48-bitový (deep color)	48	$2^{48} = 281\,474\,976\,710\,656$

Kompresa

- zmenšení velikosti souboru
- druhy komprese
 - ❖ ztrátová – vypuštění některých pixelů (např. podprahové), větší komprese zmenšuje soubor, ale snižuje kvalitu, trvalá a nevratná (JPG, MP3, MPEG, AAC)
 - ❖ bezztrátová – převod na matematický algoritmus, okolní barvy se dopočítávají, není tak účinná jako ztrátová, ale je vratná (GIF, PNG, TIFF, WMA Lossless, RealAudio Lossless, některé video kodeky – HuffuUV, Lagarith)

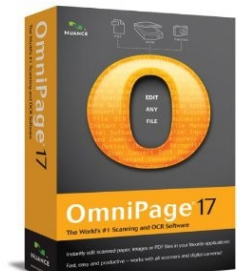
OCR

- **Optical Character Recognition**
- automatické rozpoznávání textu
 - ❖ obrazová předloha
 - ❖ analýza znaků
 - ❖ porovnání s DB (znaky, slova)
- kvalita OCR
 - ❖ přesnost rozpoznání
 - ❖ kvalita předlohy
- OCR pro národní jazyky

Nástroje - funkce

- profi i free nástroje, home verze
- ručně psané písmo – problémy
- podpora národních jazyků
- zachování layoutu a formátování
 - ❖ písmo, velikost, odstavce, obrázky
- označení bloků k rozpoznání
- ukládání jako PDF s txt vrstvou
- dávkové zpracování
- serverové verze

OmniPage



Untitled OmniPage Document 1 - OmniPage

File Edit View Format Tools Process Window Help

100% Classic View

1 - 2 - 3 Load Files Automatic Save to Files

Thumbnails Page Image

Text Editor

Style 12 Times New Roman 9 B I U

Podepsal a chlubil se
 Nezaváhal, ani když se mu osudové pero nerozepsalo. V klidu sáhl po druhém. Pak dlouho a hrde hostům ukazoval listinu se svým podpisem. A v proslovu zopakoval, že si vlastně trvá na svém. Evropská unie zůstává společem národních států, strkám unii, a tak to také nechme... Polsko zůstává zcela suverénní, prohlásil Kaczyński.
 Ale také pochválil Evropskou smlouvu je takové zlaté řeckého Evrocu. že byt msi. že ie jed

V roce 2012 konec světa nehrozi, uklidňují lidstvo potomci Mayu

Uteristi teroristé sroží zbraně

12 431 380

V roce 2012

Page	Status	Cha...	Sus...	Reje...	Wor...	Acc...	Wor...	Rec...	Ima...	Acc...	Ima...	Ima...	Ima...	Note	Rec...	Mo...	Pro...	Saved	Pen...	Boo...
1	current ...	Saved	6146	35	0	1137	100,...	9329	7,312	D:\i...	0,164	200, ...	2200...	Color						
2	Saved		2941	13	1	486	99,966	8244	3,537	D:\i...	0,147	200, ...	1696...	Color						
Total			9087	48	1	1623	99,989	8975	10,849		0,311									

Page 1 of 2

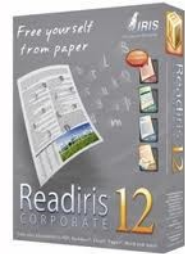
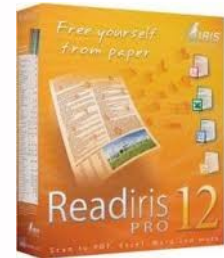
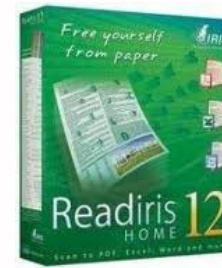
Czech Ln 009 Col 006

Omnipage

- nyní verze 18
- různé licence
 - ❖ Standard, Professional, Enterprise
- 123 jazyků
 - ❖ vč. češtiny, nemá český interface
- přesnost 99%
- propojení se zařízeními a SW
 - ❖ Kindle, MS Office
- propojení na cloud úložiště
 - ❖ Evernote a Dropbox



Readiris

The screenshot shows the Readiris 12 Corporate software interface. The main window displays a scanned newspaper article with three columns of text. The text is highlighted in yellow, indicating it has been successfully recognized by the OCR engine. The interface includes a left sidebar with various tool buttons like "Průhled OCR", "Skenovat", "Otvorit", "Tisk dokumentu", "Měnit", "Koment", "Uživatelův inkas", "Rozpoznat + Učit", "Číst", and "Převést". The top of the window shows the title bar "Readiris - OHP Program File (Readiris Corporate 12.0 Sample Czech MF (beta) 1.1.0)" and a menu bar with options like "Soubor", "Úpravy", "Nastavení", "Zobrazit", "Správa", "Účty", "Regulovat", "Nápověda". The bottom of the window shows a table with columns for "Jmeno", "Typ", "Data", "Číslo", "Stav", "Data", "Data", "Data", "Data", "Data", "Data", "Data".

PRAHA (řka) – Výdaje na zdravotní péči dosáhly v loňském roce celkem 74,1 miliardy korun, z toho 19,8 miliardy bylo vynaloženo na lůžkovou péči, 30,9 miliardy na ambulantní a 17,8 miliardy na léčiva. Ministr zdravotnictví Jan Stráský, který včera na tiskové konferenci se svými spolupracovníky hodnotil uplynulých sedm měsíců, řekl, že se podařilo mj. regulovat léčebné výkony, náklady na léky i počet zdravotních pojišťoven. Ministerstvo zdravotnictví (MZ) tak podle jeho představitelů splnilo téměř všechny úkoly, které obsahuje ministerský ozdravný Krátkodobý program.

Meziroční nárůst výdajů na léčiva v letech 1994 a 1995 byl 27,1 procenta, což je asi polovina meziročního nárůstu v předchozím období (1994/1995). Z původních 27 zdravotních pojišťoven jich dnes v ČR působí 17 a o sloučení uvažují další ústavy. Přijaté novely tzv. pojišťovacíh zákonů zpřísňují podmínky hospodaření pojišťoven. Součástí regulačních kroků MZ je i to, že nákup deseti druhů nákladné zdravotnické techniky je možný jen s jeho souhlasem.

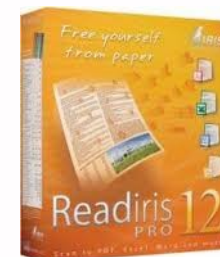
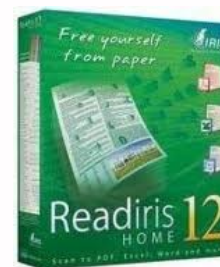
K významným opatřením v dalším období patří například připravovaná změna úhrady zdravotní péče. Od 1. ledna 1997 by měla

rychlá z počtu registrovaných pacientů zdravotního zařízení a už nikoli z vykázaných výkonů. V lůžkových zařízeních by se zavedla kombinovaná denní platba. Ke stejnému datu chystá MZ vydání nového seznamu výkonů a zřejmě i nového seznamu léčiv.

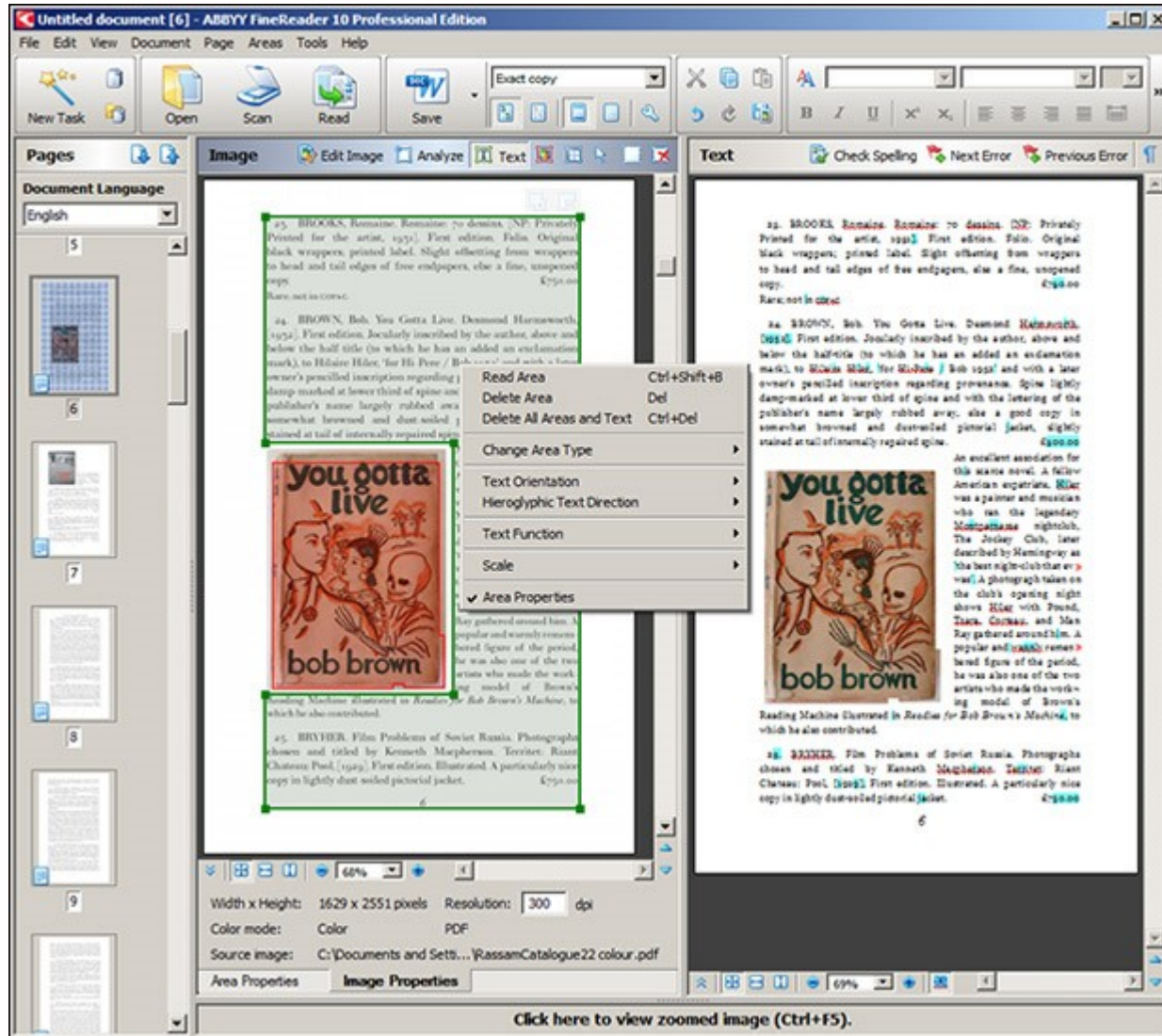
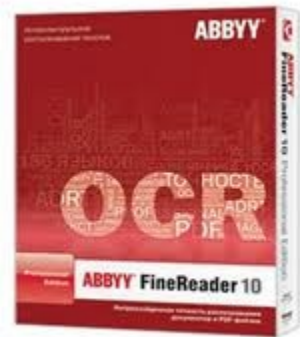
K současným protestním vystoupením zdravotníků, která budou mít spíše osvětový charakter, ministr řekl: »Děkuji těm zdravotníkům, kteří svými postoji způsobili výraznou změnu konfrontačních akcí plánovaných na tyto dny. Nepovažuji to za vítězství, ale záva-zek pro sebe.« J. Stráský uvažuje o tom, že zůstane po volbách v čele tohoto resortu.

Readiris

- nyní verze 12
- 120 jazyků (vč. češtiny)
- spolupracuje s MS Office
- konverze do PDF
- více info na Grafika.cz

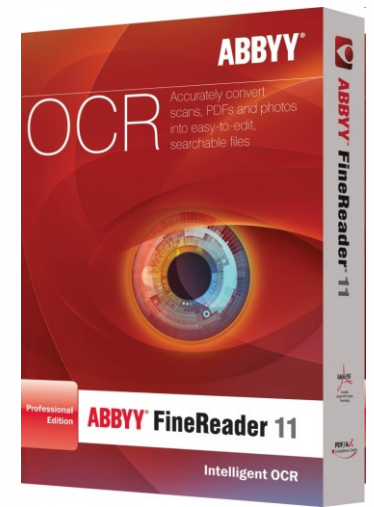


Abby Fine Reader

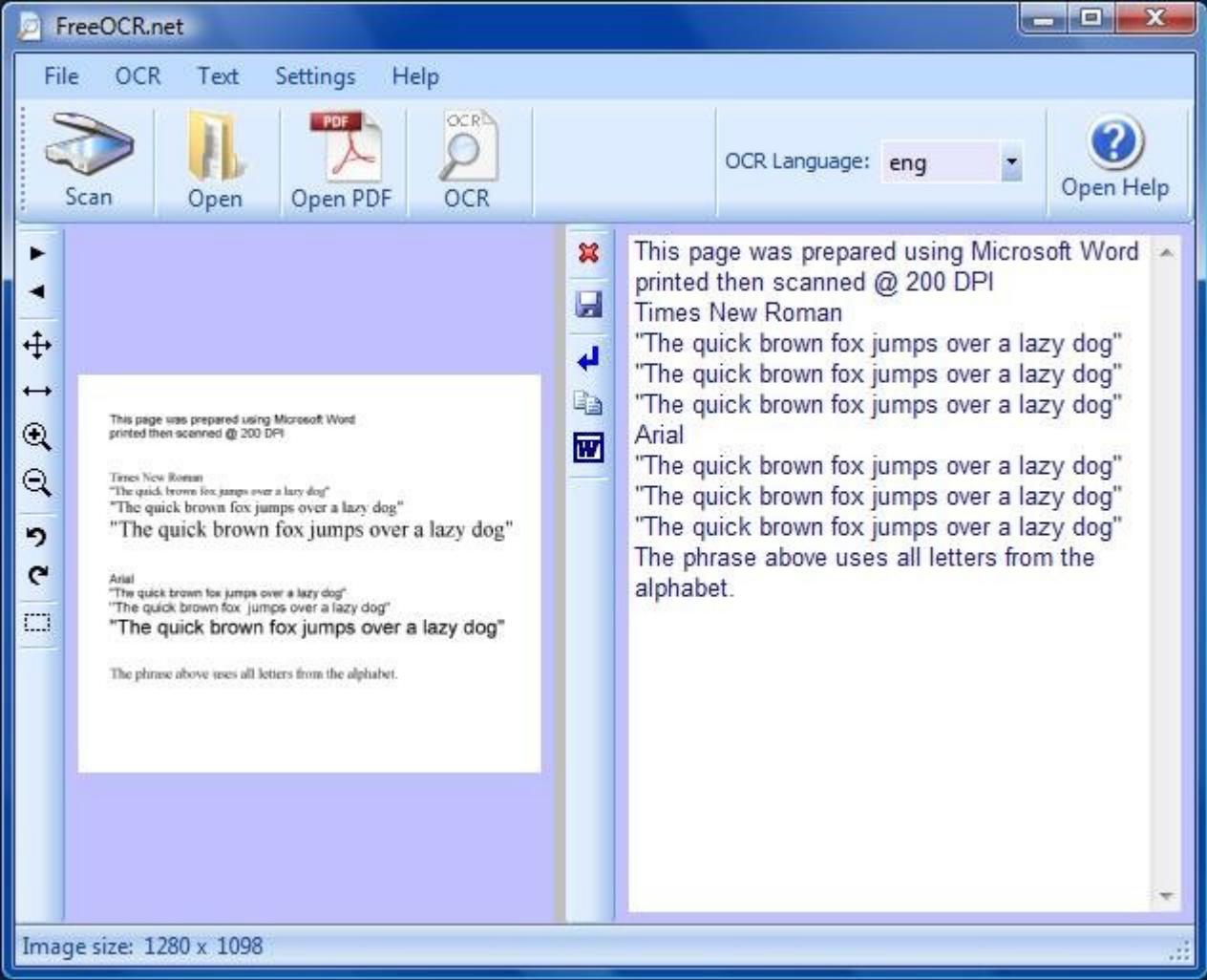


Abby Fine Reader

- verze 11
- 189 jazyků (včetně češtiny)
- stejné funkce jako konkurence
- serverová verze
- více info na Grafika.cz
- PDF Transformer 3.0
 - ❖ převod PDF do editovatelné podoby



Free OCR



FreeOCR

- OCR zdarma pro Windows
- verze 3.0
- vychází z Tesseract OCR Engine
- nemá české podklady
- horší kvalita výstupu
- solidní výsledky na podporované jazyky
 - ❖ ENG, GER, SPA, POR, NDL, ITA

Online služby

■ OnlineOCR

- ❖ 32 jazyků včetně češtiny
- ❖ dobrá kvalita, zachování v layoutu, výstupy do DOC, XLS, TXT
- ❖ omezení (odpadnou po registraci)

■ NewOCR

- ❖ 29 jazyků včetně češtiny
- ❖ posloupnost odkazů, ale nezachová layout, dobrá kvalita
- ❖ limity: obrázky 5MB, PDF (20MB)
- ❖ bez registrace

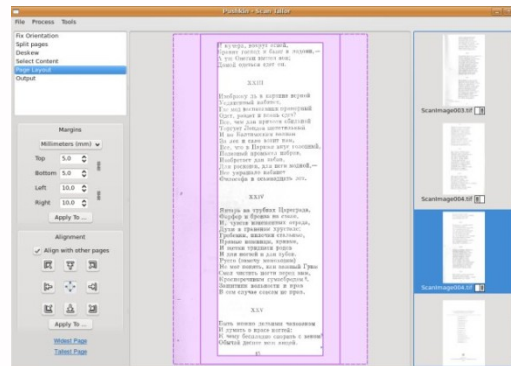
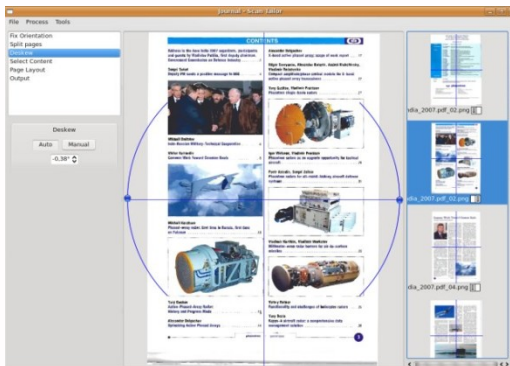
Online služby

■ Free OCR

- ❖ 29 jazyků včetně češtiny
- ❖ dobrá kvalita, asi nejlepší
- ❖ obrázky 5MB, jednostránkové PDF
- ❖ captcha

ScanTaylor

- opensource
- komplexní nástroj pro úpravu dokumentů
- otáčení, spojování, odstraňování částí stránek, OCR
- nepodporuje PDF – PDF-TO-TIFF



Články k nástrojům

- <http://extrawindows.cnews.cz/prehled-softwaru-rozpoznavani-textu-ocr-jak-na>
- <http://extrawindows.cnews.cz/prehled-softwaru-rozpoznavani-textu-ocr-jak-na?page=0,1>

Výstupní formáty

Výstupní formáty

- grafické

- ❖ JPG, TIFF, PNG, GIF, BMP

- textové

- ❖ TXT, RTF, PDF, DjVu

JPG (JPEG)

- **J**oint **P**hotographic (**E**xpert) **G**roup
- nejrozšířenější formát
- ztrátová komprese (0-100%)
- malé soubory
- vhodné na web

BMP

- bitová mapa (**bit map**)
- složen z bodů
- neumožňuje kompresi

TIFF

- **T**agged **I**nterchange **F**ile **F**ormat
- podobný BMP
- ztrátová komprese

GIF

- **G**raphics **I**nterchange **F**ormat
- v minulosti velmi populární
- využití u animovaných obrázků
- uchová max. 256 barev
- bezztrátová komprese

TXT

- Plain Text
- text bez formátování
 - ❖ pouze odstavce
- zpracuje jakýkoliv program pracující s textem

RTF

- **Rich Text Format**
- prostý text se základním formátováním
- Microsoft

PDF

- **P**ortable **D**ocument **F**ormat
- Adobe
- komerční Adobe Acrobat
- prohlížení Reader
- volně dostupné programy
 - ❖ [PDFCreator](#), [PDFill PDF Tools](#) a [další](#)
 - ❖ kancelářské balíky: [Open Office](#), MS Office 2010,...

DjVu

- konkurence PDF
- otevřený formát
- vhodný pro text, obrázky, kresby
- rozdělení do vrstev
 - ❖ výběr vhodné komprese = efektivnější komprese = malé soubory
- více info: <http://djvu.org>

Hardware pro digitalizaci

Skenery

- kvalita skenování
- zvolit vhodný skener pro konkrétní druh dokumentu!!!

Ruční a tužkové skenery



Siberian tiger
The Siberian tiger (*Panthera tigris altaica*), also known as the Amur, Manchurian, Altai, Korean, North China or Ussuri tiger is a subspecies of tiger which once ranged throughout Western, Central Asia and eastern Russia, though it is now completely confined to the Amur-Ussuri region of Primorsky Krai and Khabarovsk Krai in far eastern Siberia, where it is now protected. It is the largest of

Ruční a tužkové skenery

- problémy s kvalitou a přesností
- chyby při OCR (nejen) českých textů
- výhoda - přenosnost zařízení
- novější integrovaný disk

Plochý skener



Plochý skener

- stolní skener
- relativně kvalitní
- nízká cena
 - ❖ v multifunkcích již okolo 1000Kč
- formát A4
- velkoformátové skenery
- odrážení světla na CCD snímač
 - ❖ čím tmavší, tím menší odraz světla =
určení barvy

Rotační skener

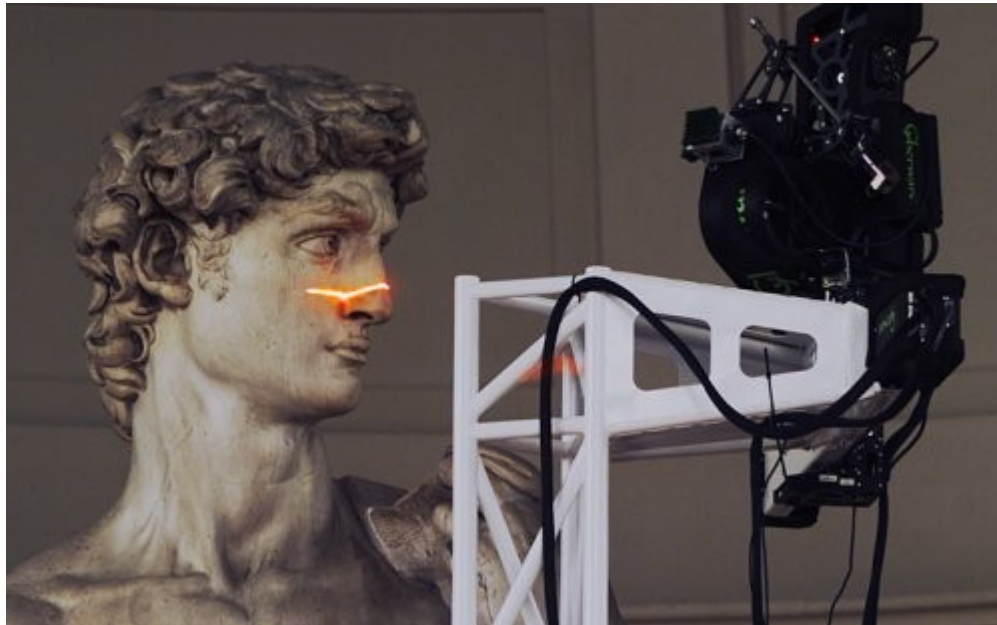


Rotační skener

- bubnový skener
- profesionální využití
- kvalitní výstup
- nelze použít na knihy
 - ❖ upínání na válec
- extrémně drahé
 - ❖ stovky tisíc až mil. Kč

3D skener

- tvorba 3D modelů



Knižní skener

- <http://www.youtube.com/watch?v=-oOXXpxzETA>



Knižní skener

- fotografické skenery
- komplexní systémy
 - ❖ vč. PC a SW
- robotické skenery
 - ❖ otáčení stránek
 - ❖ extrémně drahé

Více info o skenerech

- Skenery a skenování

- ❖ http://www.sout-prelouc.cz/stranky/polygrafie_grafika_drobek/dokumenty/maturita10/graf_18_sken.pdf

**Postup
digitalizace**

Výběr a příprava dokumentů

- vytipování dokumentů k digitalizaci
- různá kritéria výběru
- prohlédnutí a volba zařízení pro digitalizaci
 - ❖ záleží na typu dokumentu
 - ❖ digitalizace VŠKP na MU – rozřezání
 - ❖ vyčištění, narovnání listů, svorky,...
 - ❖ kvalita předlohy

Stanovení pravidel

- workflow = jednotný postup při digitalizaci, pravidla digitalizace
- uplatnění zejména v projektech a při vícenásobném skenování
- kvalita, rozlišení, komprese, formát,...
- záleží na typu dokumentu

Digitalizace

- provedení převodu
- dle pravidel
- kontrola na výstupu

Uložení do repozitáře

- zvolit vhodný archiv
- metadata
- vhodný formát
- zálohování

Využití digitalizace

- Kde se digitalizace využívá???



Použité zdroje

- <http://www.exon.cz/sluzby/digitalizace-dokumentu#postup-digitalizace>
- [http://www.posam.cz/CLANEK/12647731569780/Skenovani %28digitalizace%29 dokumentu.htm](http://www.posam.cz/CLANEK/12647731569780/Skenovani-%28digitalizace%29-dokumentu.htm)
- [http://kisk.phil.muni.cz/wiki/Digitalizace dokument%C5%AF](http://kisk.phil.muni.cz/wiki/Digitalizace-dokument%C5%AF)

Závěr



Děkuji Vám za pozornost

Martin Krčál
krcal@fss.muni.cz