

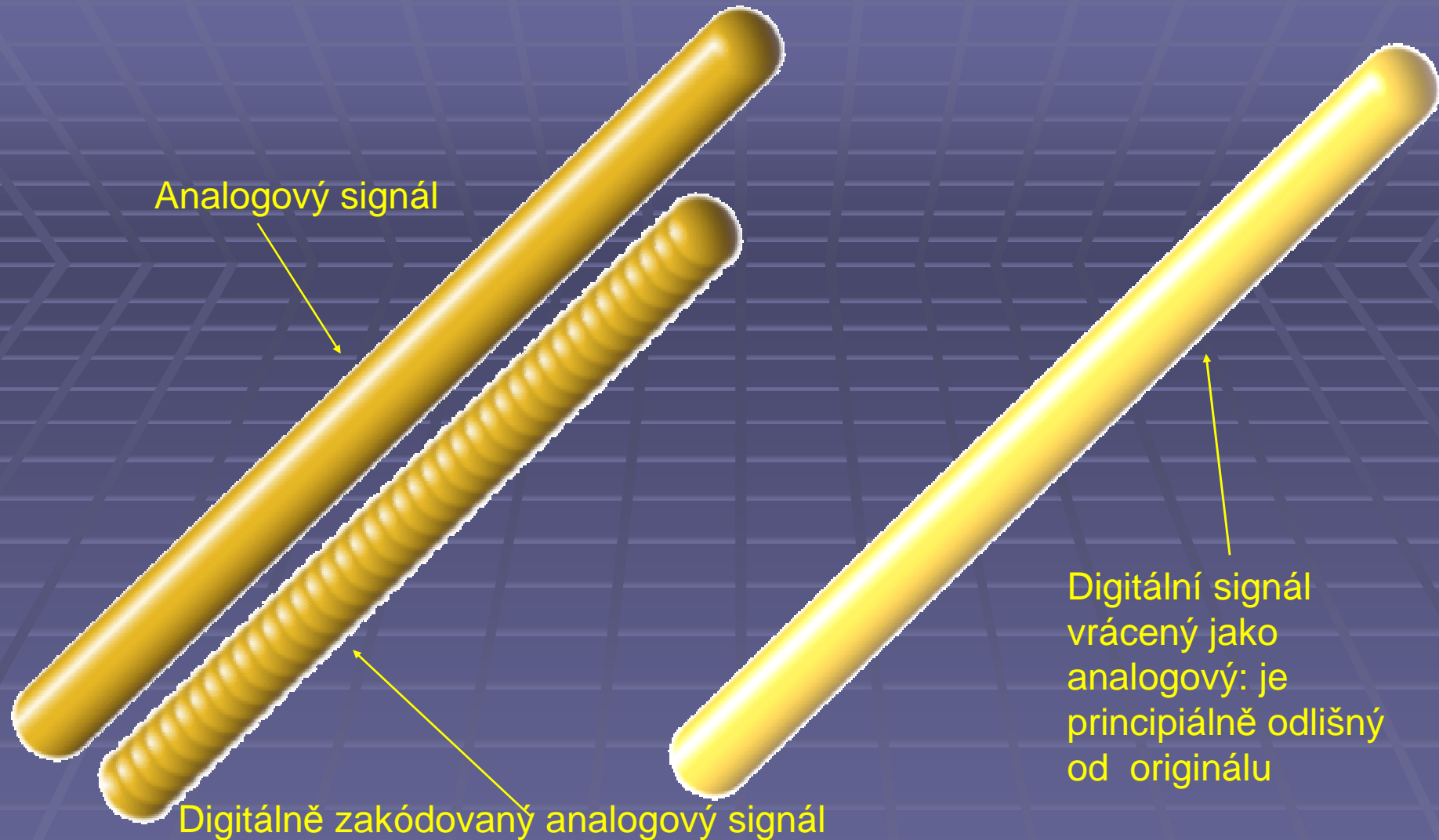
Digitalizace v českých knihovnách

Adolf Knoll

Národní knihovna České republiky

Záznam informací

- Záznam informací působících na naše základní smysly, především zrak a sluch, tzn. obrazy a zvuk
- Vždy existuje nosič informací:
 - Jsou na něm naneseny chemické látky (tisk, film)
 - Analogový záznam
 - Je uzpůsobena jeho struktura (magneticky nebo fyzicky)
 - Analogový záznam (stárne a mění se jeho kvalita, tj. zakódovaný obraz i zvuk)
 - Digitální záznam (lze ho 100% odloučit od nosiče)



Pozn.: ne všechny součásti informace byly vzaty v úvahu, aby reprezentovaly kontinuální/spojitý signál v digitalizačním procesu, ale hustota přetržitých/diskrétních částic informace by měla být s to – za součinnosti nedokonalosti našeho vnímání a jeho schopnosti předvídat – znovu vytvořit dojem spojitosti.

Rozdíly

Analogový signál

- Spojitý
- Vytváří nerozlučnou entitu s nosičem informací; tj. při kopírování se nepřenese 100% informací
- Je vnímán lidskými smysly

Digitální signál

- Přetržitý/nespojité
- Relativně nezávislý na nosiči informací; tj. při kopírování se přenese 100% informací
- Nemůže být vnímán lidskými smysly

Digitální informace musí být vždy převedena do analogové podoby, aby mohla být vnímána lidskými smysly. Jako taková je vždy určitou redukcí analogové reality.

Lidské vnímání

- Rozumíme informacím v našem vlastním kontextu, což je subjektivní a vlastní každému jedinci
- To způsobuje, že naše interpretace téže informace jsou rozdílné
- Digitalizace je prostředkem, jak zvýšit dostupnost informací; digitalizací přidáváme originálům určitou hodnotu, ale měli bychom minimalizovat náš zásah tím, že se vyhneme zbytné interpretaci originálu

Počátky

- Národní knihovna a AiP začínají spolupráci s UNESCO v r. 1992
 - Podpora programu Paměť světa: pilotní projekty, publikace na CD, školení, doporučení UNESCO
 - V jaké době: počátky běžné automatizace, na kterou nebyly finanční prostředky, nevýkonné počítače; největší nejdostupnější médium bylo CD

Od prezentací k programu

- Typickým produktem ranné éry digitalizace byla prezentace na CD-ROM (jedním z důvodů byl objem dat)
- Negativem byla provázanost prezentačního SW s daty
- Důsledkem bylo hledání cest vedoucích ke změně
 - Standardizace komplexního digitálního dokumentu (compound digital document) – velmi brzy ve světovém kontextu

Provázání dat a metadat

Východiska:

- Vývoj Internetu – ale kam se bude vyvíjet?
- Solidní platforma – SGML – ale co s ní?
- Stabilita obrazových dat – budoucí vývoj?
- Industrializace přístupu k datům?

Vývoj Internetu

- Nakolik bude webový prohlížeč hlavním nástrojem přístupu k datům obecně?
- Pokud ano:
 - Bude důležité, jaké obrazové formáty jsou doporučeny pro web, tj. zobrazitelné nativně v jeho prostředí: JPEG, GIF, později PNG, ...
 - Bude důležité, jaké mechanismy jsou užívány pro prezentaci dat (vnitřní kód): HTML, později.... (XML).....

SGML

- Doporučeno v 90. letech v literatuře pro svou flexibilitu kódování informací pro tzv. komplexní dokumenty (compound documents), ale prakticky pro digitalizaci?
- **PROTOŽE:**
 - Obtížná výroba (popis, struktura a odkazování)
 - Obtížná prezentace jdoucí mimo internetové prohlížeče
- Z těchto důvodů zvolena v r. 1996 vlastní cesta využití SGML, tj. tzv. DOBM formát (kombinace formálního a obsahového značkování/markupu)

DOBM

- Obecná specifikace dána DTD
- Konkrétní specifikace dána zasunovatelným *.sgm definičním souborem pro:
 - Rukopisy a staré tisky
 - Periodika

Formát byl využíván do r. 2002 v obou oblastech

Přístup typu DOBM jako doporučení UNESCO v r. 1999 pro program Paměť světa

Formáty metadat jako mezinárodní a národní standardy

Rukopisy

- 1996 – DOBM
- 2002 – `masterx.xsd` – (projekt EU, TEI P.4 MASTER schéma + mapování struktur + technická metadata o barvě z Data Dictionary for Still Digital Images a DIG.35)
- 2009 – projekt EU ENRICH, TEI P.5 schéma pro popis a strukturování rukopisů (tzv. `enrich.dtd`) jako mezinárodní standard

Novodobé dokumenty

- 1999 – DOBM
- 2002 – vlastní XML DTD pro:
 - Periodika
 - Monografie
 - (Sbírkové předměty)

Stabilita obrazových dat

- Překvapivě velká odolnost formátu JPEG (v čase ještě upevnil své postavení)
- Problematika archivních formátů (JPEG, ...)
- Problematika formátů pro zpřístupnění:
 - Parametrizovaný **JPEG**, GIF, PNG
 - Moderní newebové formáty: DjVu (nasazen po rozsáhlých testech wavelet a MRC formátů + na určitou dobu MrSID pro mapy)

Industrializace přístupu k datům

- **Digitální knihovny**
 - Manuscriptorium
 - Kramerius
- **Portály**
 - JIB
 - TEL
 - EUROPEANA
- **Distribuované digitální knihovny**
 - Manuscriptorium 2

Digitální knihovny

- Databáze + datové úložiště (databanka), tj. nikoli prosté webové stránky/seznamy
- Řešení:
 - Knihovní katalog + aplikace na prohlížení dat
 - Specializovaná digitální knihovna, rozvíjející funkce vlastní digitalizovaným dokumentům a vycházející vstříc příslušné cílové skupině uživatelů
- Nezbytné minimálně:
 - Identifikační popis
 - Strukturální mapa dokumentu

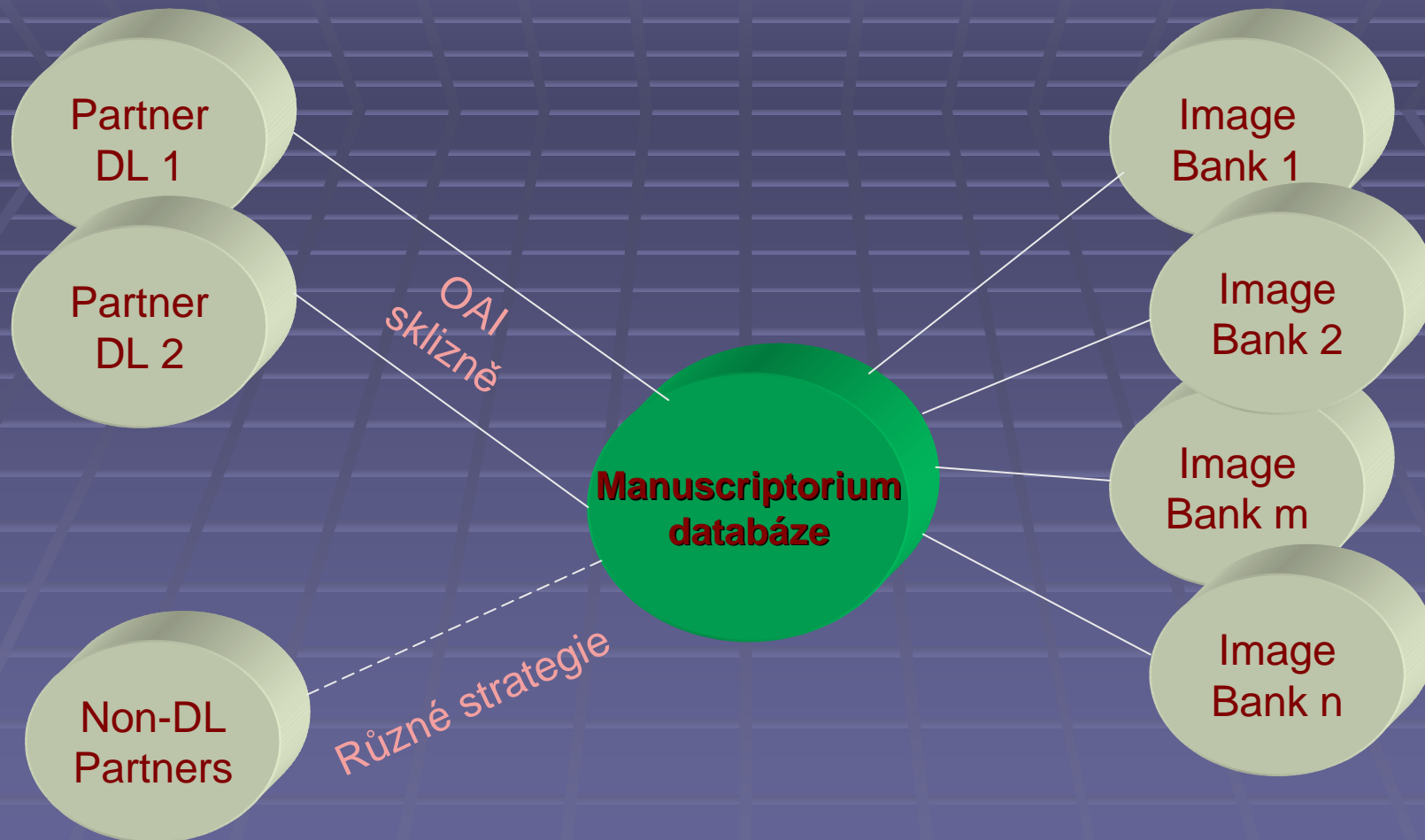
Portály

- Optimalizace vyhledávání z více zdrojů v jediném rozhraní
- Komunikace s aplikacemi vč. digitálních knihoven většinou na elementární úrovni popisných metadat (DC)
- Od Z39:50 k OAI-PMH, a tím k XML pro sdílení dat
- Pro prohlížení dat navigace do původních aplikací někdy kromě strukturálně jednoduchých dat (audio nebo video soubor), spustitelných v nativním prostředí portálu a/nebo prostředky počítače uživatele

Distribuovaná digitální knihovna

- Vychází z principu jednotného rozhraní, které využívá dat různých aplikací v prostředí Internet
- Re-use (paralelní) využití dat reálných digitálních knihoven a úložišť
- Problematika přemapování struktur a strategií poskytování rozšířených metadat
- Výsledkem je bezešvý homogenizovaný přístup (seamless access) k různorodým zdrojům

Bezešvý přístup k distribuovaným zdrojům



Data jsou volána do interface Manuscriptoria

Personalizace pro uživatele

- Vytvoření individuálního prostředí přímo v digitální knihovně:
 - Tvorba individuálních sbírek
 - Statické
 - Dynamické
 - Tvorba virtuálních dokumentů z digitálních objektů Manuscriptoria nebo z Internetu

Personalizace pro přispěvatele

- Podpora výroby dat přímo v prostředí digitální knihovny:
 - M-TOOL On-line pro práci s TEI P.5 pro popis a strukturální mapování
 - Manuscriptorium pro kandidáty: on-line prostředí pro testování a upload hotových XML souborů

Budoucnost

- Zvýšení produkce dat (národní financování, strukturální fondy, partnerství Google)
- Spolupráce s velkými hráči při zpřístupnění (nadmárodní portály, komerční služby)
- Manuscriptorium:
 - Růst agregace dat, tj. subagregace pro Europeanu
 - Plné texty
 - Ontologie, tezaury, databáze třetích stran
 - Další personalizace