

MASARYKOVA UNIVERZITA
FAKULTA INFORMATIKY



Cambridge Digital Library

PV070 - ZÁVEREČNÁ ESEJ

Martin Šalata

Brno, 25.11.2017

Základné údaje

Názov: Cambridge Digital Library
Nositeľ projektu: University of Cambridge
URL: <https://cudl.lib.cam.ac.uk/>

Charakteristika

Cambridgeská digitálna knižnica je projekt Cambridgeskej univerzitnej knižnice zameraný na sprístupnenie unikátnych kolekcí vo svojej zbierke širokej verejnosti prostredníctvom internetu.

Ciele projektu

Projekt digitálnej knižnice si stanovuje za cieľ obohacovať spoločnosť pomocou vzdelania a výskumu na najvyššej úrovni. Snaží sa to dosiahnuť vytýčením niekoľkých základných vízií [1]:

1. Prístup – Snaha sprístupniť unikátne kolekcie čo najširšej verejnosti
2. Spolupráca – Vytváraním partnerstiev s regionálnymi, národnými, ale aj medzinárodnými inštitúciami zdieľať znalosti a rozširovať spoločný dosah
3. Zapojenie – Snaha nielen sprístupniť obsah, ale aj podporiť iniciatívu o zapojenie sa verejnosti do procesu rozvoja obsahu
4. Inovácia – Zabezpečenie dlhodobého zachovania dedičstva, aby digitalizovaný obsah ostal relevantný a použiteľný aj napriek neustálym zmenám v technológiách
5. Inšpirácia – Mnohé z digitalizovaných kolekcí sa zameriavajú na inšpiratívne témy a osobnosti, aby projekt dokázal vzbudiť záujem o vedomosti, poznanie a výskum.

História

Prvá fáza tvorby digitálnej knižnice, zameraná predovšetkým na vybudovanie technickej infraštruktúry, prebehla v rokoch 2010 až 2014. Sponzorovaná bola organizáciou The Polonsky Foundation, ktorá sa zameriava na podporu rozvoja digitálnych knižníc a digitálnu knižnicu v Cambridge podporila sumou 1,5 milióna libier [1].

Prvá fáza projektu bola zameraná na digitalizáciu diel z oblasti vedy a viery. Ako prvé boli z oblasti viery zdigitalizované niektoré z islamských manuskriptov, najmä vďaka podpore organizácie The Islamic Manuscript Association. Z prvých vedeckých diel digitalizovaných Cambridgeskou knižnicou je najvýznamnejšia zbierka Newtonových prác a diárov, o ktorej rozvoj sa najviac zaslúžila nezisková organizácia Jisc prostredníctvom svojho programu v rokoch 2011 až 2013 [2].

Najvýznamnejšie zbierky

Darwinove manuskripty

Diela, zápisky a poznámky Charlesa Darwina sú postupne digitalizované a zverejňované zároveň v Cambridgeskej digitálnej knižnici a prostredníctvom tzv. Darwin Manuscripts Project¹. V súčasnosti obsahuje viac ako 30000 strán zdigitalizovaných manuskriptov z rokov 1835-1882.

Kolekcia *Creation of the Origin of Species* obsahuje prvé náznaky Darwinovho záujmu odhaliť pôvod živočíšnych druhov. Táto kolekcia zhromažďuje teoretické poznámky a množstvo esejí, ktoré Darwin napísal v priebehu dvoch desaťročí (1837-1859). Eseje predstavujú dôležitú súčasť zbierky, pretože zachytávajú myšlienkové procesy Darwina počas formovania základov teórie prirodzeného výberu.

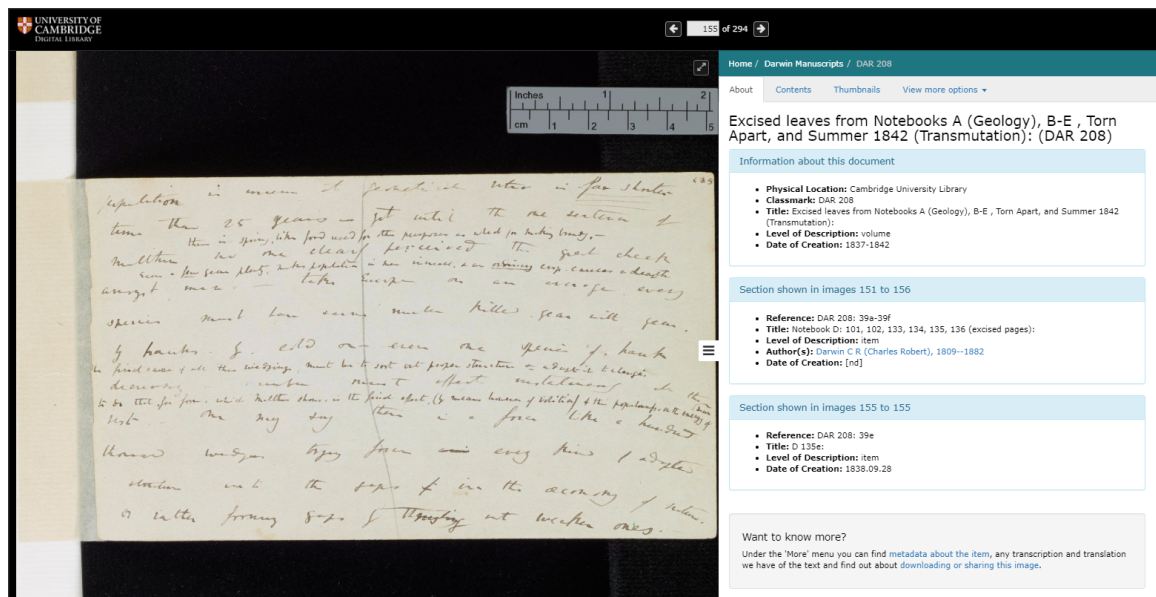
Špeciálne významné sú poznámky z prvého "zošitu premeny"²(Zošit B, 1837-1838), v ktorom sa Darwin po prvý krát pokúsil sformulovať teóriu o evolúcii. Neskôr v zošitoch D a E (1838-1839) začal formulovať teóriu prirodzeného výberu.

Creation of the Origin of Species taktiež zachytáva postup návrhov výsledného diela *Pôvod druhov*, keď ako prvú z návrhových esejí napísal v roku 1842 a už v tomto veľmi skorom návrhu (*Pôvod druhov* bol publikovaný v roku 1859) zachytil štruktúru a popis základnej problematiky. Bolo to v tejto eseji, keď po prvý krát použil pojem *prirodzený výber*.

Druhá významná kolekcia s názvom *Darwin's Evidence* sa zameriava na praktický aspekt Darwinovho výskumu. Obsahuje dokumentáciu k rozsiahlym experimentom vykonaným v období po dokončení *Pôvodu druhov*, ktoré však slúžili aj na podporu tvrdení tohoto diela. Výskumy sa zameriavali napríklad na komplexné metódy šľachtenia, fyziológiu rastlín alebo fertilizáciu.

1. Projekt Amerického prírodovedeckého múzea. <https://www.amnh.org/our-research/darwin-manuscripts-project>

2. Transmutation notebook, pomenovanie pre poznámkové zošity zachytávajúce myšlienky o evolúcii



Obr. 1: Ukážka webového rozhrania (na obrázku je strana z Darwinovho denníka)

Newtonove diela

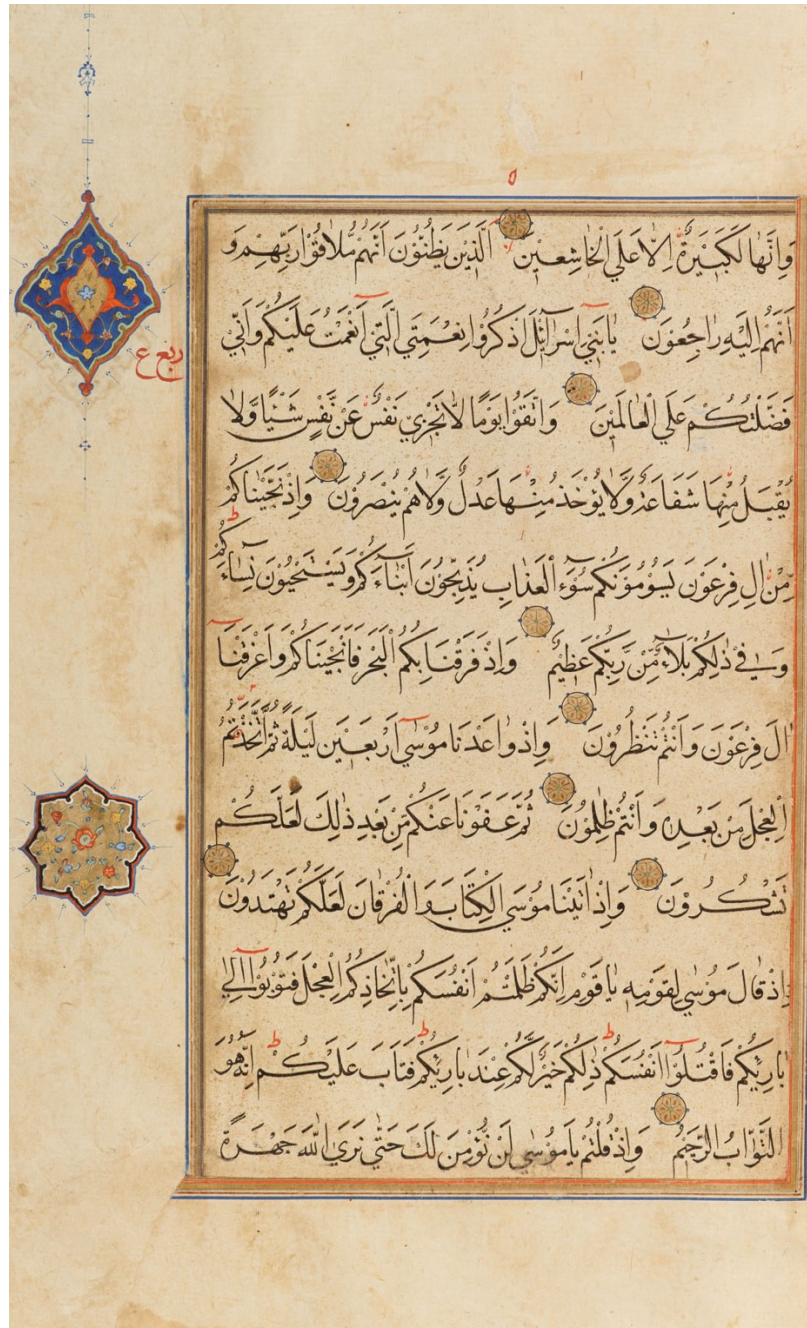
Je to jedna z prvých zdigitalizovaných zbierok z oblasti vedy. Zbierka Newtonových diel je najväčšia a najvýznamnejšia zbierka jeho vedeckých prác na svete. Okrem vedeckých prác obsahuje aj rôzne diáre a listy [4]. Spolu s digitalizovanými dielami nahrala knižnica aj niekoľko videí vysvetľujúcich Newtonovu prácu a približujúcich niektoré z manuskriptov.

Islamské manuskripty

Ich zbierka začala v knižnici vznikať už okolo roku 1630 a dnes obsahuje viac ako 5000 diel. Okrem iného bola z tejto zbierky zdigitalizovaná aj plná verzia ilustrovaného Koránu zachovaná z prelomu prvého a druhého tisícročia [3]. Rovnako bola zdigitalizovaná aj celá zbierka manuskriptov z prvých štyroch storočí Islamu.

Mapy

Technologicky oveľa náročnejšie na spracovanie ako bežné texty a knihy sú nepochybne veľkoformátové mapy. Cambridgeská knižnica obsahuje viac ako 1,3 milióna máp a v zdigitalizovanej podobe je možné nájsť mapy datované od 15. storočia



Obr. 2: Ukážka zdigitalizovanej strany Koránu

až po súčasnosť. Najväčšia zdigitalizovaná mapa v knižnici (zároveň aj najväčšia zdigitalizovaná položka v knižnici) má rozmery 405 x 259 cm a jedná sa o mapu distribúcie ryže v časti Burmy (dnešné Mjanmarsko) [5].

Iné

Okrem textov, malieb a máp sú súčasťou digitálnej knižnice aj rôzne ďalšie zaujímavé objekty, ako napríklad glóbusy, staroveké hudobné nástroje, kusy oblečenia a podobne [6].

Prístup

Základným prístupovým bodom k informáciám v Cambridgeskej digitálnej knižnici je webové rozhranie [6]. Toto rozhranie ponúka ako aj jednoduché, tak aj komplexné možnosti vyhľadávania v knižnici [7] (vyhľadávanie je možné spresniť napríklad podľa rozmedzia rokov vzniku diela, miesta, kde vzniklo, autora a podobne).

V súčasnosti je webové rozhranie jedinou možnosťou, ako k digitalizovaným dielam pristupovať. Do budúca je v pláne vytvoriť API na prístup k metadátam diel, bližšie informácie k tomuto API však v aktuálne nie sú k dispozícii. Cambridgeská univerzitná knižnica, ktorej diela digitálna knižnica digitalizuje, však svoje API pre prístup k metadátam má [8].

Štandardy

Knižnica nemá zjednotené používanie metadátových formátov. Pre rôzne druhy diel sa používajú štandardy TEI, METS, and MODS [9]. TEI je štandard pre reprezentáciu textových diel v digitálnej podobe. Hlavným cieľom je strojovo-čitateľná podoba. Cambridgeská digitálna knižnica ho používa vo formáte:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>...</fileDesc>
    <encodingDesc>...</encodingDesc>
    <profileDesc>...</profileDesc>
  </teiHeader>
  <facsimile>...</facsimile>
  <text>...</text>
</TEI>
```

V hlavičke, `fileDesc` popisuje metadáta diela (bibliografický záznam). `encodingDesc` a `profileDesc` popisujú formát uložených dát a polia `facsimile` a `text` popisujú samotný text (jednotlivé strany zdigitalizovaného dokumentu a príslušné odkazy do webového rozhrania). Môžeme si všimnúť, že týmto formátom sú popisované aj texty, ktoré boli zdigitalizované, ale sú dostupné len vo forme fotografií/skenov bez prepisu skenovaného textu. V tomto prípade dátová časť TEI súboru obsahuje len odkazy na jednotlivé preskenované strany. Týmto spôsobom je zdigitalizovaná napríklad dizertačná práca Stephena Hawkinga [10].

METS je XML schéma určená k vyjadreniu hierarchickej štruktúry objektov digitálnych knižníc a príslušných metadát.

MODS je XML schéma určená k vyjadreniu bibliografických dát. Cambridgeská digitálna knižnica používa formáty METS a MODS spoločne na definíciu zdigitalizovaných diel. Výsledný tvar schémy je nasledovný:

```
<mets:metsHdr/>
  <mets:dmdSec ID="DMD-1" ADMID="AMD-1">
    ...
  </mets:dmdSec>
  <mets:amdSec ID="AMD-1">...</mets:amdSec>
  <mets:amdSec ID="AMD-2">...</mets:amdSec>
  <mets:amdSec ID="AMD-3">...</mets:amdSec>
  <mets:fileSec>...</mets:fileSec>
  <mets:structMap TYPE="PHYSICAL">
    ...
  </mets:structMap>
  <mets:structMap TYPE="LOGICAL">
    ...
  </mets:structMap>
</mets:METS>
```

V tomto súbore tagy `mets:dmdSec` s `ID="AMD-X"` obsahujú informácie o licencovaní, konkrétne práva spojené s metadátami, sťahovaním obrázkov a zobrazovaním obrázkov. Tag `mets:fileSec` popisuje súbory, ktoré spolu vytvárajú dané dielo a tagy `mets:structMap` definujú samotné dielo. Tag `mets:dmdSec` obsahuje bibliografické informácie vo formáte MODS:

```
<mets:dmdSec ID="DMD-1" ADMID="AMD-1">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods xmlns:mods="..." ID="...">
        <titleInfo xmlns:xsi="..." xmlns="...">...</titleInfo>
        <abstract>...</abstract>
        <name xmlns:xsi="..." xmlns="..." type="personal"
          usage="primary">
```

```
    ...
    </name>
    <name xmlns:xsi="..." xmlns="...">...</name>
  </mods:mods>
</mets:xmlData>
</mets:mdWrap>
</mets:dmdSec>
```

Týmto formátom je popísaná väčšina obrazových diel v digitálnej knižnici, napríklad mapy [11].

Záver

Cambridgeská digitálna knižnica je projekt v pomerne ranom štádiu vývoja. Množstvo dostupných zdigitalizovaných diel nie je veľké. Knižnica taktiež neposkytuje žiadne API na prístup k zdigitalizovaným metadátam (okrem samotného webového rozhrania). Rovnako neexistuje veľa dostupných informácií o použitých technológiách a o procese digitalizácie diel.

Projekt digitálnej knižnice pravidelne zverejňuje krátke správy s najnovšími prírastkami do zdigitalizovanej zbierky. Podrobnejšie informácie o aktuálnom stave knižnice (celkový počet zdigitalizovaných diel, plán digitalizácie v najbližšej budúcnosti) však chýbajú.

Celkovo tento projekt pôsobí dojmom začínajúceho projektu, ktorý si bude vyžadovať veľké množstvo prostriedkov a úsilia na dosiahnutie požadovanej kvality. Projekt však má podľa môjho názoru budúcnosť, a to najmä vďaka dostupným prostriedkom a množstvu unikátnych a zaujímavých zbierok a diel Cambridgeskej univerzitnej knižnice, ktoré by stálo zato zdigitalizovať a poskytnúť širokej verejnosti.

Metadata

```
<?xml version="1.0" encoding="UTF-8"?>

<metadata
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<dc:title>Cambridge digital library</dc:title>
<dc:creator>Martin Šalata</dc:creator>
<dc:subject>Digitálna knižnica</dc:subject>
<dc:description>Esej zameraná na popis vzniku, rozvoja
a technologického zabezpečenia fungovania Cambridskej
digitálnej knižnice</dc:description>
<dc:date>2017-11-25</dc:date>
<dc:type>Text</dc:type>
<dc:format>application/pdf</dc:format>
<dc:language>sk</dc:language>
<dc:rights>public</dc:rights>

</metadata>
```

Bibliografia

- [1] Anne Jarvis. *CDL: Introducing the Cambridge Digital Library*. 2011. URL: <https://cudl.lib.cam.ac.uk/about/> (cit. 2017-11-22).
- [2] Nathalie Southall. *JISC Grant Funding 11/10: eContent Programme*. 2013. URL: https://www.webarchive.org.uk/wayback/archive/20140615012610/http://www.jisc.ac.uk/fundingopportunities/funding_calls/2010/09/grant1110.aspx (cit. 2017-11-22).
- [3] *CDL: Islamic Manuscripts*. 2015. URL: <https://cudl.lib.cam.ac.uk/collections/islamic> (cit. 2017-11-22).
- [4] *CDL: Newton Papers*. 2015. URL: <https://cudl.lib.cam.ac.uk/collections/newton> (cit. 2017-11-22).
- [5] *CDL: Maps*. 2015. URL: <https://cudl.lib.cam.ac.uk/collections/maps> (cit. 2017-11-22).
- [6] *CDL: Browse our collections*. 2017. URL: <https://cudl.lib.cam.ac.uk/collections/> (cit. 2017-11-22).
- [7] *CDL: Search*. 2017. URL: <https://cudl.lib.cam.ac.uk/search/> (cit. 2017-11-22).
- [8] *CUL: Newton Search Service*. URL: <http://www.lib.cam.ac.uk/api/docs/newtonSearch.htm> (cit. 2017-11-22).
- [9] *CDL: Help: FAQ*. URL: <https://cudl.lib.cam.ac.uk/help/> (cit. 2017-11-22).
- [10] *Properties of expanding universes / Stephen Hawking, TEI metadata*. URL: <https://services.cudl.lib.cam.ac.uk/v1/metadata/tei/MS-PHD-05437> (cit. 2017-11-22).
- [11] *Map of Cantabrigia*. URL: <https://services.cudl.lib.cam.ac.uk/v1/metadata/mods/PR-ATLAS-00004-00057-00003> (cit. 2017-11-22).