SPRÍSTUPŇOVANIE A UCHOVÁVANIE VEDECKÝCH DÁT

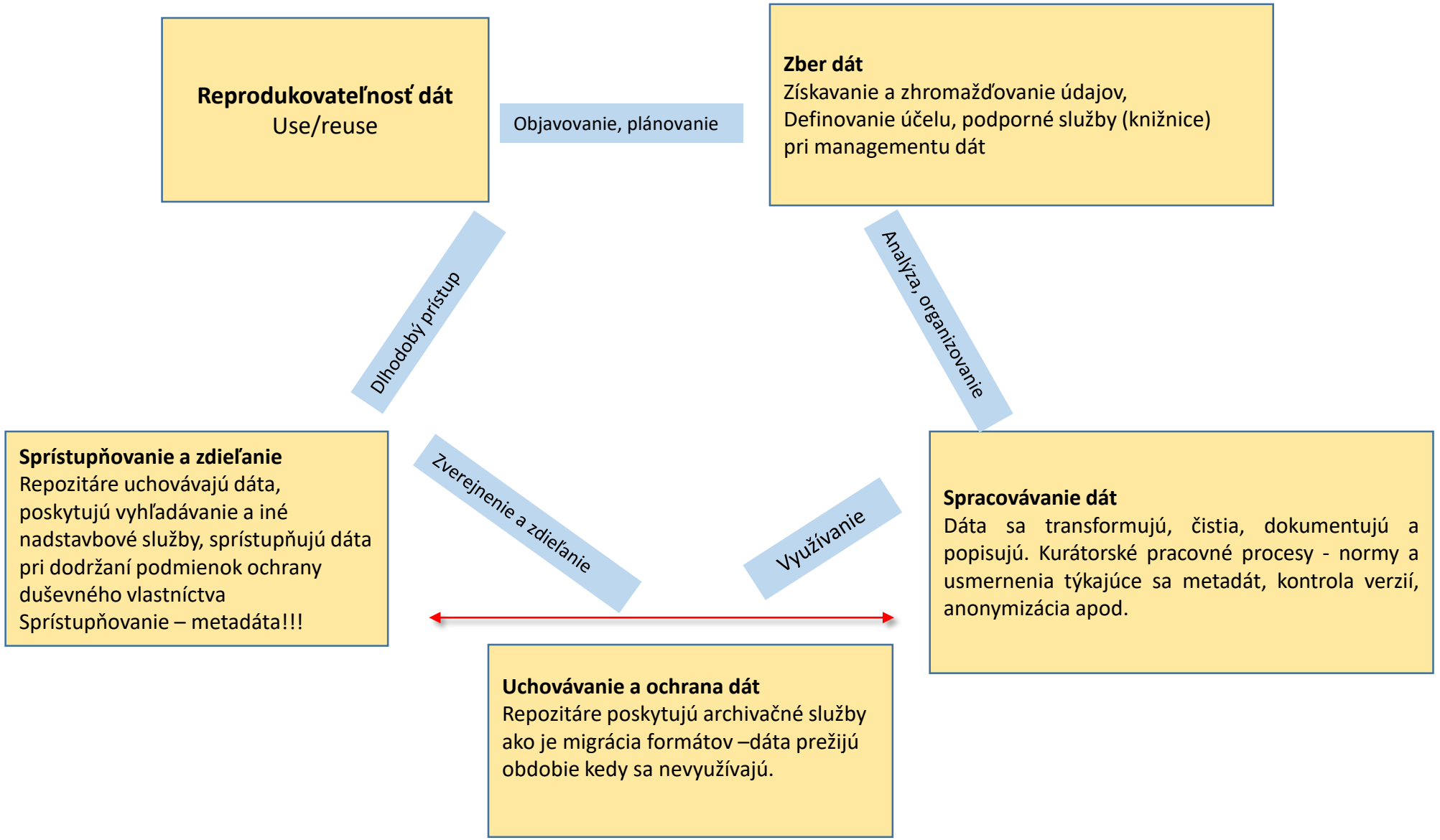
Jitka Dobbersteinová, vzdelávací program Otvorená veda v praxi – Otvorené vedecké dáta, 17.10. – 19. 10. 2023

EURÓPSKA ÚNIA
Európske štrukturálne a investičné fondy
OP Integrovaná infraštruktúra 2014 – 2020

MINISTERSTVO DOPRAVY A VÝSTAVBY SLOVENSKEJ REPUBLIKY

MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY

CENTRUM VEDECKO-TECHNICKÝCH INFORMÁCIÍ SR



Legenda:
Žlté polia: kurátorský cyklus dát
Modré šípky: životný cyklus managementu vedeckých dát

Kredit: https://www.dcc.ac.uk

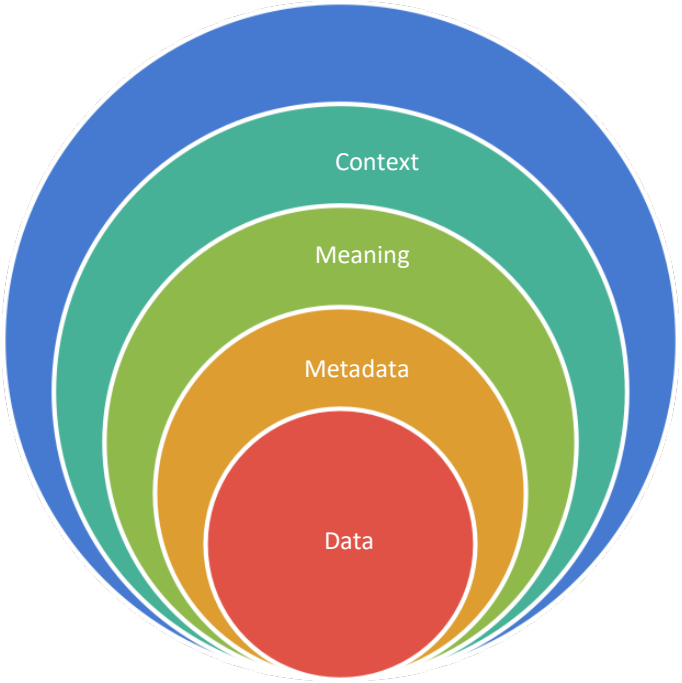
Získanie, analýza

1

Spracovanie

DÁTA vs METADÁTA?

- Digitálny objekt môže odvodzovať veľkú čas svojho významu z atribútov, ktoré nie sú samotnou jeho súčasťou
  - Kontext – okolnosti, za ktorých vznikol
  - Cesta – ako vznikol a dostal sa tam, kde je
  - To platí najmä pre dáta



- Metadáta môžu niesť viac významu ako samotné dáta
- **Vieme vôbec jednoznačne definovať metadáta?**

AKO ZABEZPEČÍTE ZROZUMITEĽNOSŤ DÁT PRE INÝCH VEDCOV...?

Opis použitých štandardov a metód pomôže pochopiť iným, ako ste realizovali projekt

(dokumentácia – kontextové informácie, na úrovni dát, na úrovni štúdie, nástroje; readme súbor)

AKO BUDETE DÁTA DOKUMENTOVAŤ, OPISOVAŤ?

Metadát je veľa – ktoré budú kľúčové pre projekt?

Metadáta môžu byť:

Opisné metadáta – informácie o obsahu a kontexte dát

Napríklad: názov, tvorca, kľúčové slová predmetu a opis (abstrakt)

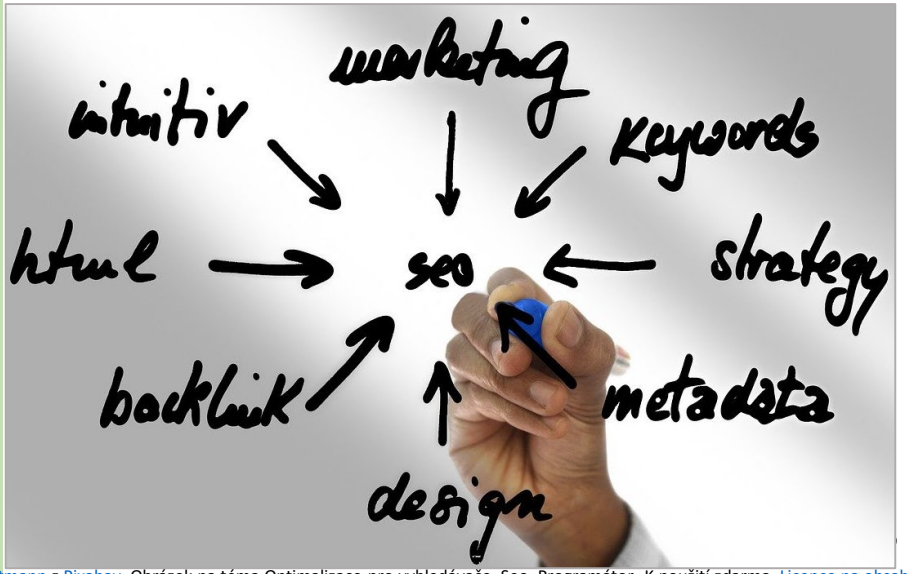
Štrukturálne metadáta – opisujú fyzickú štruktúru zložených dát

Napríklad: použitý fotoaparát, clona, expozícia, formát súboru a vzťah k iným údajom alebo súborom

Administratívne metadáta – informácie používané na manažment dát

Napríklad: kedy a ako boli vytvorené, kto k nim môže pristupovať, softvér potrebný na ich použitie a autorské práva

Zdroj: https://guides.lib.unc.edu/metadata/definition



METADÁTA = DÁTA O DÁTACH

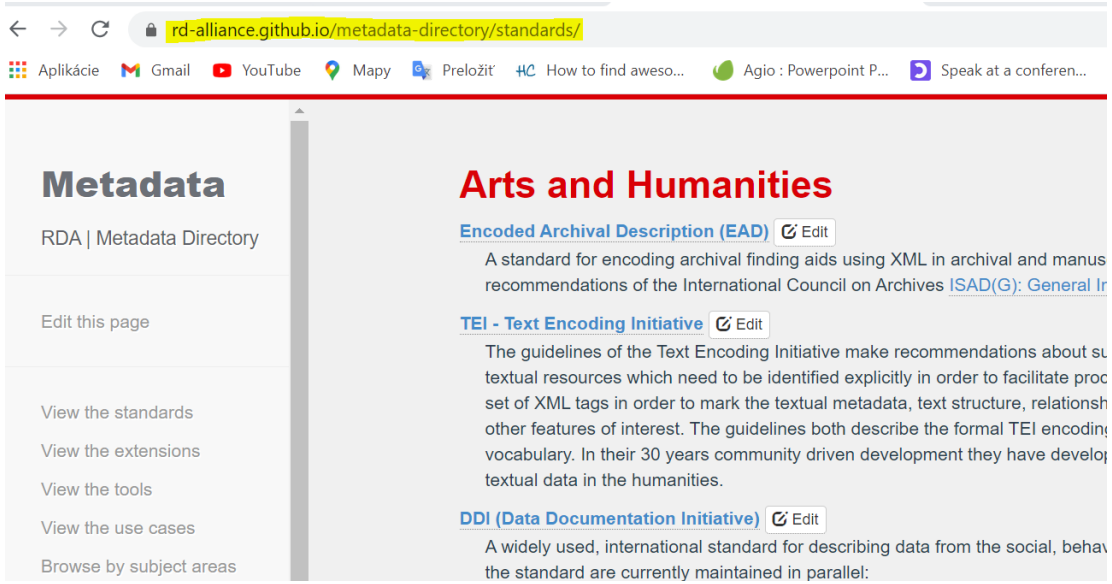
- Popisujú **charakteristiky** objektu/zdroja : obsah, formát, lokalita, prístupové práva...
- Popisujú **fyzické** alebo **digitálne** objekty
- **Forma:**
  - **voľný text** (readme file)
  - **štruktúrovaná a štandardizovaná** forma (MARC, DublinCore...)
- **Výskumné dáta** : nájditeľnosť, citovanosť a opätovné použitie
- **Spojené s dátami**, ktoré popisujú:
  - spolu v dátovom súbore
  - separátny súbor
  - záznam v katalógu/repozitári

Metadátové štandardy

https://rd-alliance.github.io/metadata-directory/standards

Štandardy:

- Dáta sú rovnako štruktúrované
- Štandard podľa disciplíny: **odborové metadátové štandardy,** napr. DDI, Darwin Core, EML, VRA Core...
- Repozitáre si môžu určovať konkrétne štandardy
- ISO, Dublin Core Metadata Initiative....



MARC 21

Dublin Core™

https://www.dcc.ac.uk/guidance/standards/metadata

Sample Dublin Core metadata for an image

Below is an example of Dublin Core metadata for this photograph of Howard Walter Florey. Generic examples for the Dublin Core metadata schema are available from the Dublin Core website.

- **Title** = Howard Walter Florey
- **Author/Creator** = Not known
- **Subject and Keywords** = The Australian National University photographs
- **Subject and Keywords** = Nobel Laureates
- **Description** = This photograph was taken at the 1948 Easter Conference, which considered the early planning of the Australian National University, at the Institute of Anatomy building. A keen amateur photographer, Florey carries his camera over his shoulder.
- **Publisher** = The Australian National University Digital Collections
- **Date** = 1948
- **Type** = image
- **Format** = image/tif
- **Resource identifier** = `http://hdl.handle.net/1885/7491`
- **Resource identifier** = ANUA225-405-3
- **Source** = ANU Archives
- **Rights** = This image is provided for research purposes only and must not be reproduced without the prior permission of the Archives Program, Australian National University

https://libguides.anu.edu.au/c.php?g=881167&p=6358464

Metadátové schémy

- **štruktúra** metadát
- **špecifikuje elementy** = povinné, voliteľné, opakovateľné
- vychádza zo špecifickej komunity (knihovníci - MARC), popisuje špecifický formát al. domény (ISO 19115)
- môže špecifikovať: obsahové pravidlá (formáty, riadené slovníky), syntax (napr. XML)

Example of MARC 21 record

```
=LDR  00000nam\a2200000\a\4500
=005  20111115115722.0
=008  111126s2008\\\\ii\a\\\\\b\\\\001\0\eng\\
=020  \\$a9780073046235
=040  \\$aBD-DhIUB$cBD-DhIUB
=082  00$a332.1068$222
=100  1\$aRose, Peter S.
=245  10$aBank management & financial services /$cPeter S. Rose, Sylvia C. Hudgins.
=246  3\$aBank management and financial services
=250  \\$a7th ed.
=260  \\$aNew York :$bMacmillan,$cc2008.
=300  \\$axxv, 722 p. :$bill. ;$c25 cm.
=504  \\$aIncludes bibliographical references and index.
=650  \0$aFinancial institutions$zUnited States.
=650  \0$aBank management$zUnited States.
=700  1\$aHudgins, Sylvia Conway,$d1956-
```

```
<head profile="http://dublincore.org/documents/dcq-html/">
  <title>Dublin Core</title>
  <link rel="schema.DC"      href="http://purl.org/dc/elements/1.1/" />
  <link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
  <meta name="DC.format"        scheme="DCTERMS.IMT"      content="text/html" />
  <meta name="DC.type"          scheme="DCTERMS.DCMIType" content="Text" />
  <meta name="DC.title"         content="Dublin Core" />
  <meta name="DC.publisher"     content="Jimmy Wales" />
  <meta name="DC.subject"       content="Dublin Core Metadaten-Elemente, Anwendungen" />
  <meta name="DC.creator"       content="Björn G. Kulms" />
  <meta name="DCTERMS.license"  scheme="DCTERMS.URI" content="http://www.gnu.org/copyleft/fdl.html" />
  <meta name="DCTERMS.rightsHolder" content="Wikimedia Foundation Inc." />
  <meta name="DCTERMS.modified" scheme="DCTERMS.W3CDTF" content="2006-03-08" />
</head>
```

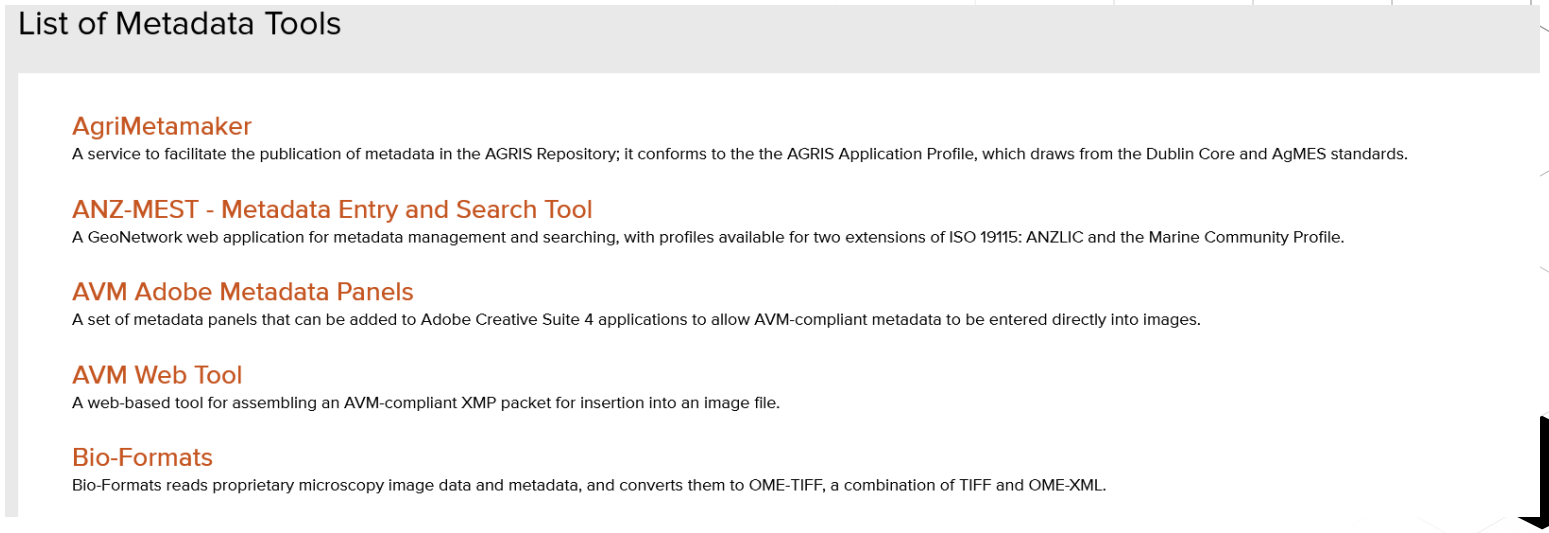
dc.contributor.author	SCHÜTZE, Fabian
dc.coverage.spatial	USA
dc.coverage.temporal	1959-2013
dc.date.accessioned	2019-10-16T15:43:18Z
dc.date.available	2019-10-16T15:43:18Z
dc.date.created	2013-2014
dc.date.issued	2014
dc.identifier.other	EUI_ResData_00001_ECO
dc.identifier.uri	http://hdl.handle.net/1814/64586
dc.description	1 data file. Dataset elaborated from Robert Shiller stock market database, 1992-2013; and Federal Reserve Economic Data (FRED) on aggregate consumption.
dc.description.abstract	Dataset of adjusted monthly data for aggregate U.S. consumption expenditures and equity returns, 1959-2013, sourced from Robert Shiller stock market database, 1992-2013; and Federal Reserve Economic Data (FRED) on aggregate consumption - generated in the context of a research project illustrating an econometric technique, Generalised Method of Moments (GMM).
dc.format	Excel file
dc.format.mimetype	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
dc.language.iso	en
dc.publisher	European University Institute, ECO
dc.relation.ispartofseries	EUI Research Data
dc.relation.ispartofseries	2014
dc.relation.ispartofseries	Department of Economics

dc.rights	info:eu-repo/semantics/openAccess
dc.rights.uri	http://creativecommons.org/licenses/by/4.0/
dc.subject	Statistical data
dc.subject	Consumption
dc.subject	Equity prices
dc.subject	E20
dc.subject	E21
dc.subject	G12
dc.subject.classification	YP-CS
dc.subject.ddc	317.3
dc.subject.lcsh	Consumption (Economics) -- United States -- Statistics -- Databases
dc.subject.lcsh	Stocks -- Prices -- United States
dc.title	Dataset of adjusted monthly data for aggregate U.S. consumption expenditures and equity returns, 1959-2013
dc.type	Dataset
dc.identifier.doi	10.2870/342675
eui.subscribe.skip	true
dc.rights.license	Creative Commons Attribution 4.0 International

B2 | Ben Kok

	A	B	C	D	E	F
1	ID	author	date	subject.age	subject.gender	DataFile
2	sample 1	Ben Kok	-10-2015		male	FLProj_int_subj1.txt
3	sample 2		12-10-2015		male	FLProj_int_subj2.txt
4	sample 3		12-10-2015		female	FLProj_int_subj23.txt
5	sample 4		12-10-2015		female	FLProj_int_subj4.txt

https://www.dcc.ac.uk/guidance/standards/metadata/tools



zdieľanie a zverejnenie dát

Princípy FAIR a vedecké dáta

AKO BUDETE DÁTA DOKUMENTOVAŤ, OPISOVAŤ?

Metadát je veľa – ktoré budú kľúčové pre váš projekt?

RIADENÉ SLOVNÍKY

Príklady a odkazy na riadené slovníky
https://guides.lib.utexas.edu/metadata-basics/controlled-vocabs

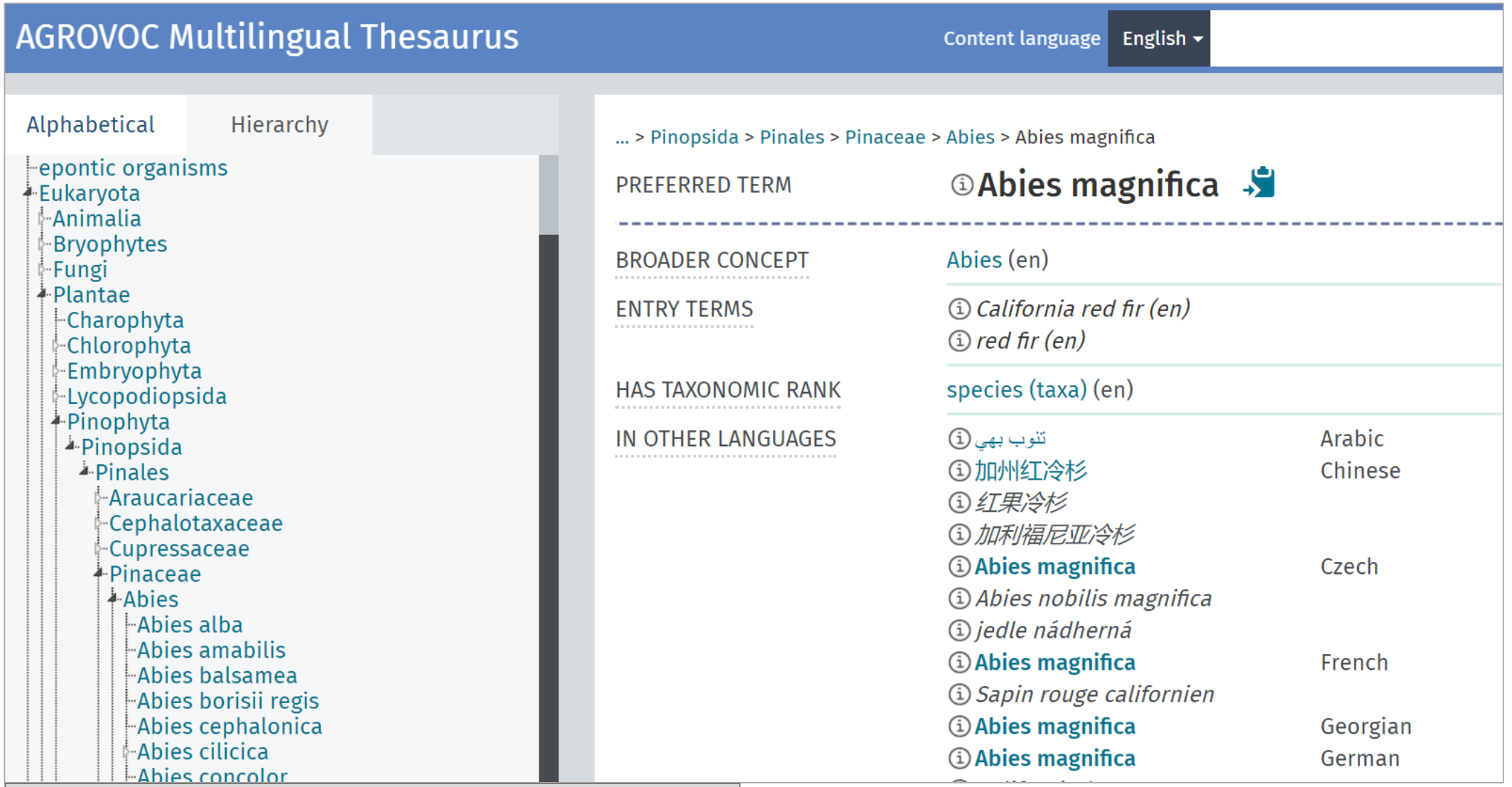
Library of Congress Authorities for subject headings
https://authorities.loc.gov/

ICD-11: International Classification of Diseases 11th Revision
The global standard for diagnostic health information
https://www.who.int/standards/classifications/classification-of-diseases

The Astronomy Thesaurus
https://www.mso.anu.edu.au/library/thesaurus/

Art & Architecture Thesaurus® Online
Getty Research Institute
http://www.getty.edu/research/tools/vocabularies/aat/

RIADENÉ SLOVNÍKY - PRÍKLAD



https://agrovoc.fao.org/browse/agrovoc/en/page/c_11?clang=sk

2

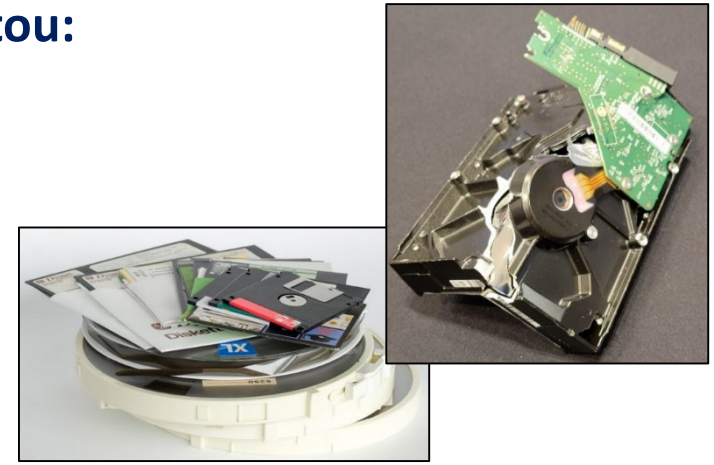
Uchovávanie Ochrana

DLHODOBÉ UCHOVÁVANIE VÝSKUMNÝCH DÁT PO UKONČENÍ PROJEKTU

Ktoré z množstva získaných dát majú byť **uchovávané dlhodobo?**

Posúdiť hodnotu svojich dát, dlhodobo uchovávať len dáta s vysokou hodnotou:

- všetky podkladové dáta súvisiace s publikáciou
- dáta, ktorých zber je náročný alebo nákladný
- dáta, ktoré sa nedajú nahradiť/znovu vygenerovať



KTORÉ DÁTA MAJÚ BYŤ **ZNIČENÉ/VYMAZANÉ? AKO?**

- biologický materiál, ktorý sa časom znehodnotí
- dôverné údaje alebo osobné údaje, ktoré nemožno anonymizovať

Nástroj DCC 'Five steps to decide what data to keep: a checklist for appraising research data'

AKO DLHO PO UKONČENÍ PROJEKTU DÁTA UCHOVÁVAŤ?

- minimálne obdobie pre dáta, ktoré sú podkladom pre publikácie
- dodržiavať právne predpisy, nariadenia platné pre váš typ dát, pravidlá či usmernenia inštitúcie...

- Zabezpečenie dát v dôveryhodnom repozitári/úložisku
- Uchovávanie dát počas a po projekte
- Plány dlhodobej ochrany, archivácie (veľkosť dát, ako dlho, zničenie dát, finančné náklady, stratégie proti strate dát)

Výber riešenia pre uchovávanie dát https://www.re3data.org/

Typy riešení		
inštitucionálny repozitár	**publikácie, ale aj dáta**	
disciplinárny repozitár		**https://www.re3data.org/**
inštitucionálne úložisko	**úložiská prispôsobené na uchovávanie dátových súborov**	
národná/nadnárodná infraštruktúra		**národné repozitáre, Zenodo, World Data System**
komerčné úložiská	**úložiská prispôsobené na uchovávanie dátových súborov**	**Google Drive, Figshare, Dropbox...**

Bezpečnosť, formáty dát

- Nekompatibilita
- Univerzálné formáty: ASCII, Unicode
- „Bezpečné“ formáty
- Zálohovanie

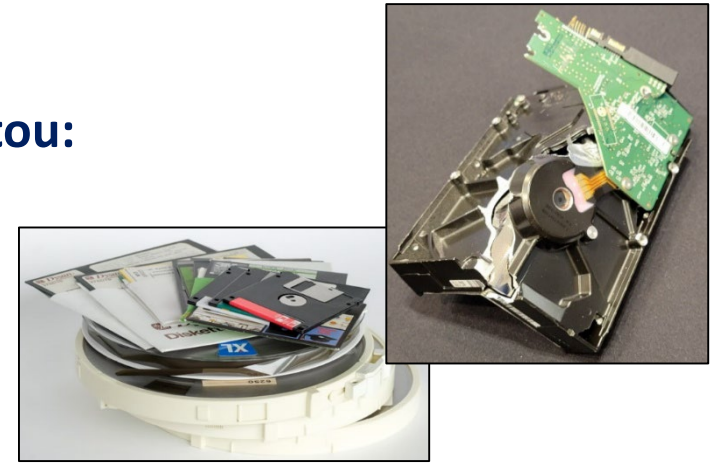
Type	Preferred format(s)	Non-preferred format(s)
Text documents	• PDF/A (.pdf) • ODT (.odt)	• Microsoft Word (.doc) • Office Open XML (.docx) • Rich Text File (.rtf) • PDF other than PDF/A (.pdf)
Markup language	• XML (.xml) • HTML (.html) • Related files: .css, .xslt, .js, .es	• SGML (.sgml) • Markdown (.md)
Programming languages	• MATLAB • NetCDF • Text-Fabric	
Spreadsheets	• ODS (.ods) • CSV (.csv)	• Microsoft Excel (.xls) • Office Open XML Workbook (.xlsx) • PDF/A (.pdf)
Databases	• SQL (.sql) • SIARD (.siard) • CSV (.csv)	• Microsoft Access (.mdb, .accdb) • dBase (.dbf) • HDF5 (.hdf5, .he5, .h5)
Raster images	• JPEG (.jpg, .jpeg) • TIFF (.tif, .tiff) • PNG (.png) • JPEG 2000 (.jp2) • DICOM (.dcm)	
Vector images	• SVG (.svg)	• Adobe Illustrator (.ai) • EPS (.eps) • WMF/EMF (.wmf, .emf)

DANS Dutch national centre of expertise and repository for research data

Uchovávanie dát

DLHODOBÉ UCHOVÁVANIE VÝSKUMNÝCH DÁT PO UKONČENÍ PROJEKTU

Ktoré z množstva získaných dát majú byť **uchovávané dlhodobo?**



Posúdiť hodnotu svojich dát, dlhodobo uchovávať len dáta s vysokou hodnotou:

- všetky podkladové dáta súvisiace s publikáciou
- dáta, ktorých zber je náročný alebo nákladný
- dáta, ktoré sa nedajú nahradiť/znovu vygenerovať

Nástroj DCC 'Five steps to decide what data to keep: a checklist for appraising research data,
Nástroj na vyhodnotenie nákladov na uchovávanie dát Data Management costing tool and checklist (UK Data Service)

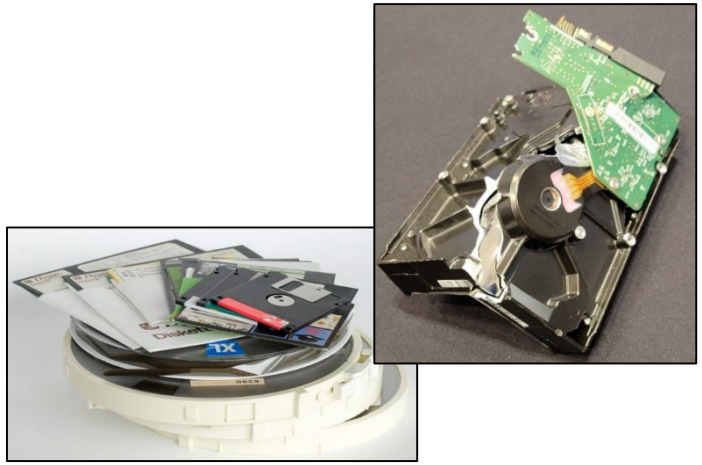
AKO DLHO PO UKONČENÍ PROJEKTU DÁTA UCHOVÁVAŤ?

- minimálne obdobie pre dáta, ktoré sú podkladom pre publikácie
- dodržiavať právne predpisy, nariadenia platné pre váš typ dát, pravidlá či usmernenia inštitúcie...

Uchovávanie dát

KTORÉ DÁTA MAJÚ BYŤ **ZNIČENÉ/VYMAZANÉ? AKO?**

- biologický materiál, ktorý sa časom znehodnotí
- dôverné údaje alebo osobné údaje, ktoré nemožno anonymizovať
- usb kľúče
- papier, optické disky



Softvér na vymazanie súborov z pevných diskov: BCWipe, WipeFile, DeleteOnClick a Eraser pre platformy Windows; a Permanent Eraser pre platformy MacOS.

Zdroj: https://www.linkedin.com/pulse/why-your-data-must-fair-david-crosswell

Ďakujem
Otázky?