Kvalitativní analýza výzkumných dat Univerzity Karlovy v Praze

Qualitative analysis of the research data at Charles University in Prague

Eliška Pavlásková

Ústav informačních studií a knihovnictví (Univerzita Karlova v Praze), Ústav dějin Univerzity Karlovy a archiv Univerzity Karlovy

Recenzenti:
Prof. PhDr. Jela Steinerová, PhD.
Mgr. Ilona Trtíková

Abstrakt

Článek je zaměřen na aktuální tématiku výzkumných dat. Je založen na výsledcích kvalitativního výzkumu provedeného na Univerzitě Karlově v Praze. Cílem výzkumu bylo odvození základních vlastností výzkumných dat na základě analýzy textů disertačních prací ze sbírek Univerzity Karlovy v Praze. Výzkum byl motivován zejména potřebou dlouhodobého uložení těchto dat a využívá tedy primárně teoretické koncepty a terminologii z této oblasti. Data (disertační práce) byla analyzována metodou obsahové analýzy a metodou zakotvené teorie. V tomto článku jsou prezentovány výsledky kvalitativní analýzy.

Klíčová slova: *dlouhodobé uchovávání, výzkumná data*

Abstract

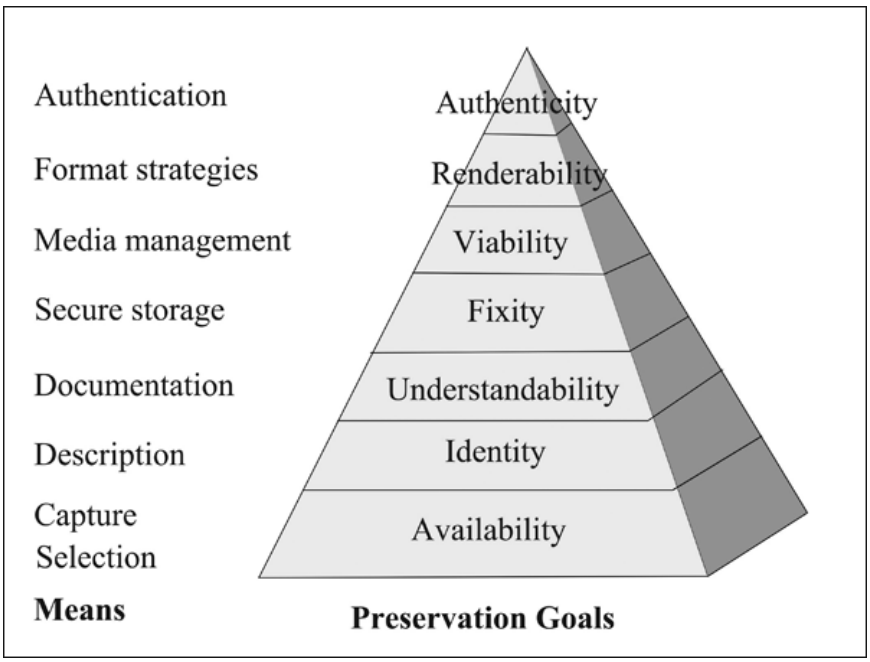
The text focuses on the current topic of research data. It presents results of qualitative research at Charles University in Prague. The aim of the research was to establish main characteristics of research data by content analysis of the sample of dissertations originated on Charles University in Prague. Research was motivated by the need of long-term preservation of this data and therefore uses primarily theoretical concepts and terminology from this area. Collected data (dissertation theses) were analysed by the methods of content analysis and grounded theory. Results of qualitative analysis are presented in this article.

Keywords: *Long term preservation, research data*

1. Úvod

Struktura vědecké komunikace se s nástupem digitálních technologií významně proměnila, a tato změna přirozeně zasáhla i univerzity. Nové technologie umožňují nové formy vzdělávání, výzkumu i sdílení vědeckých výsledků či výukových materiálů. Informační zdroje nejen že vznikají za přispění digitálních technologií, ale v drtivé většině případů jsou – alespoň v některé z fází svého životního cyklu – také vyjádřeny v digitální formě. Tato změna zasáhla prakticky všechny složky vědecké komunikace. Jedním z nejvýznamnějších produktů technologické změny je vznik a zejména výrazné rozšiřování množiny výzkumných dat. Britský Economic and Social Research Council (instituce stojící u vzniku UK Data Archive) vnímá data jako hlavní přínos ekonomického a sociálního výzkumu.[1]

Priscilla Caplan[2] navrhuje pyramidové schéma (viz obr. 1) popisující oblasti, na něž by se instituce při práci s digitálními objekty měla soustředit. V základu pyramidy je dostupnost objektu. Instituce by měla být schopna shromáždit objekty, které jsou pro ni z dlouhodobého hlediska relevantní a také již v této první fázi identifikovat jejich význačné vlastnosti s ohledem dlouhodobého uchovávání. V případě výzkumných dat se – zejména z institucionálního pohledu – jedná o netriviální otázku, která vyžaduje rozsáhlý a komplexní výzkum.



Obr. 1 - Pyramidové schéma dlouhodobé archivace

Dlouhodobé uchovávání vědeckých dat klade na instituce zodpovědné za jejich archivaci poměrně vysoké a v současné době navíc ještě ne zcela prozkoumané nároky. Motivací níže

[1] ECONOMIC AND SOCIAL RESEARCH COUNCIL. Research data policy. In: *Economic and social research council* [online]. ESRC [online]. 2015 [cit. 2016-01-27]. Dostupné z: http://www.esrc.ac.uk/funding/guidance-for-grant-holders/research-data-policy/

[2] CAPLAN, Priscilla. Chapter 1: What is digital preservation? *Library technology reports* [online]. 2008, **44**(2) [cit. 2016-03-18]. ISSN 0024-2586. [cit. 2014-12-27]. DOI: 10.5860/ltr.44n2. Dostupné z: https://journals.ala.org/ltr/article/view/4224/4809

popsaného výzkumu byla zejména potřeba zmapování situace v rámci konkrétní instituce (v tomto případě se jednalo o Univerzitu Karlovu v Praze), tak i potřeba ověření poznatků získaných zejména ze zahraničních zdrojů, jejich aplikace na české prostředí a zasazení do kontextu dlouhodobého uchovávání. Terminologicky text vychází především z českého překladu normy ČSN ISO 14721[3] a terminologie používané v rámci standardu PREMIS.[4]

Výzkum byl proveden v rámci připravované disertační práce, která také bude obsahovat jeho kompletní výsledky. Výzkumné metody a způsob výběru výzkumného vzorku jsou popsány v kapitole 3. Zde vzhledem k omezenému rozsahu článku uvádím pouze výsledky kvalitativní analýzy vybraného vzorku. Výsledky kvalitativní analýzy jsou obsaženy v kapitole 4 tohoto článku. Analýza byla provedena s ohledem na předběžnou identifikaci faktorů ovlivňujících dlouhodobé uložení výzkumných dat. Část provedeného výzkumu zaměřená zejména na strukturu výzkumných dat používaných v rámci zkoumaného vzorku byla prezentována na konferenci Archivy, knihovny a muzea v digitální době[5]. Tento článek a uvedená prezentace se shodují pouze v oblasti teoretických východisek výzkumu.

2. Teoretická východiska výzkumu – definice a vlastnosti výzkumných dat

ČSN ISO 14721 definuje data jako: „opakovaně interpretovatelná vyjádření informací ve formalizované podobě vhodné pro komunikaci, interpretaci nebo zpracování; mezi příklady dat patří posloupnost bitů, tabulka s čísly, znaky na stránce, nahrávka zvuků pořízená mluvčím nebo vzorek měsíční horniny.“[6] Tato definice je obecně přijímaná v komunitě zabývající se dlouhodobým uložením digitálních objektů. Pro oblast vysokých škol a univerzit (respektive pro ty z nich, které se věnují vědecké a výzkumné činnosti) má význam zejména problematika podskupiny výzkumných dat, tedy dat, které jsou podkladem pro vědeckou práci. Vlastnosti dat se do určité míry odvíjejí od způsobu jejich sběru. Dělení dle zdroje dat bylo navrženo v rámci dokumentu *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, který byl vypracován National Science Board.[7] Z tohoto dokumentu vychází i dělení výzkumných dat, které ve svých pracích navrhuje Borgmanová.[8,9] Vyděluje čtyři skupiny dat

[3] ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. 97 s.

[4] CAPLAN, Priscilla. *Understanding PREMIS* [online]. Washington: Library of Congres, 2009 [cit. 2016-02-13]. Dostupné z: http://www.loc.gov/standards/premis/understanding-premis.pdf

[5] PAVLÁSKOVÁ, Eliška. *Výzkumná data na Univerzitě Karlově v Praze*. 16. konference Archivy, knihovny, muzea v digitálním světě 2015.

[6] ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. s.20

[7] NATIONAL SCIENCE BOARD. *Long-lived digital data collections: Enabling research and education in the 21st century* [online]. 2005. Arlington: National Science Foundation [cit. 2016-02-13]. Dostupné z: http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

[8] BORGMAN, Christine L. The conundrum of sharing research data. *Journal of the American society for information science and technology* [online]. 2012, **63**(6), s. 1059-1078 [cit. 2016-02-13]. DOI: 10.1002/asi.22634. ISSN 15322882. Dostupné z: http://doi.wiley.com/10.1002/asi.22634

[9] BORGMAN, Christine L. *Big data, little data, no data: scholarship in the networked world*. Cambridge: The MIT

(NSB navrhuje pouze tři), přičemž je třeba poznamenat, že toto dělení považuje za aplikovatelné pouze v oblasti přírodních a sociálních věd nikoli v medicíně a v humanitní oblasti. Další zpřesnění tohoto dělení přináší Hendl, který detailněji definuje rozdíl mezi observačními a experimentálními daty.[10]

Observační data jsou data pocházející z pozorování. Objekty jsou sledovány a jsou měřeny proměnné, ale zároveň proměnné nejsou ovlivňovány. Klasickým příkladem observačních dat jsou meteorologická data sledující proměnné ovlivňující nebo vytvářející počasí – například data o objemu srážek. V sociálních vědách se může jednat o rozhovory nebo o etnografická pozorování.

Komputační (výpočetní) data jsou výsledkem počítačového modelování, simulací či workflow.

Experimentální data jsou výsledkem procesu, během kterého je cíleně ovlivňována závislá proměnná. Experiment je prováděn v kontrolovaném prostředí.

Záznamy jsou čtvrtou kategorií, kterou ke skupinám definovaným NSB doplnila Borgmanová. Z hlediska výzkumných dat se jedná o záznamy související s daty samotnými. Mohou to být terénní poznámky, historické záznamy nebo ručně psané deníky. Borgmanová je definuje neobyčejně široce jako „téměř jakýkoli fenomén či lidská aktivita, která může být využita jako data pro výzkum.[11]“

3. Zkoumaný vzorek a metody výzkumu

Mým záměrem bylo vytvořit předběžný průřezový obraz využívání výzkumných dat v rámci univerzity jako celku. Z toho důvodu jsem si jako zdroj dat vybrala sbírku závěrečných disertačních prací. Jedná se o sbírku zahrnující všechny vědecké obory, která však zároveň není zkreslena rozdílnou publikační praxí mezi přírodními a sociálními či humanitními vědami. Zároveň se jedná o prokazatelně vědecké publikační výsledky. Volba vzorku však byla motivována i pragmaticky – jedná se o ucelenou a snadno dosažitelnou sbírku. Konkrétně jsem pracovala se vzorkem 400 prací, které byly vybrány na základě kvótního výběru tak, aby bylo zachováno poměrové zastoupení jednotlivých fakult. Procento prací z dané fakulty zastoupených ve vzorku tedy odpovídá procentu zastoupenému v populaci. V rámci fakulty byly práce vybírány náhodně, a to za využití generátoru náhodných čísel. Jednalo se o práce z let 2011–2015 (zhruba do června) napsané v češtině, slovenštině nebo angličtině. Vzorek nebyl vytvořen na základě žádného tematického či oborového omezení.

Na základě textu disertací jsem identifikovala, s jakými daty autor pracoval a jak je využíval. Zejména práce z přírodních věd obvykle obsahují kapitolu věnovanou metodice a práci s daty, která obsahuje relevantní informace. Tento přístup neumožňoval prozkoumat jednotlivé typy dat skutečně do hloubky, ale bylo díky němu možné provést plošnou analýzu neomezenou na konkrétní obor.

Press, 2015. xxv, 383 stran. ISBN 978-0-262-02856-1.

[10] HENDL, Jan. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Vyd. 2., opr. Praha: Portál, 2006. 583 s. ISBN 80-7367-123-9.

[11] BORGMAN, Christine L. *Big data, little data, no data: scholarship in the networked world*. Cambridge: The MIT Press, 2015. xxv, 383 stran. ISBN 978-0-262-02856-1.

Za záznamovou jednotku byla považována konkrétní disertační práce. Pouze v případě, že práce očividně obsahovala více než jeden datový set a tyto datové sety byly výrazně odlišného charakteru a byly na sobě vzájemně nezávislé, byl za záznamovou jednotku považován konkrétní set. U záznamových jednotek byly sledovány následující parametry:

1. Původ dat – rozlišována byla data interní, tedy produkovaná a v ideálním případě i archivovaná v rámci Univerzity Karlovy v Praze, a data externí, tedy data pocházející z vnějších zdrojů.
2. Zařazení dat do formátové skupiny – rozlišována byla data obrazová, zvuková, audiovizuální, data textového charakteru, data numerického charakteru, data softwarového charakteru a analogové zdroje.
3. Zdroj dat – bylo použito výše uvedené dělení na observační, experimentální a komputační data a záznamy.
4. Informace o využitých softwarových aplikacích.
5. Zařazení do vědní oblasti – bylo použito dělení na přírodní, sociální a humanitní vědy.

Na základě obsahové analýzy textu práce jsem provedla kvalitativní analýzu za využití metody zakotvené teorie,[12] která je definována jako „teorie induktivně odvozená ze zkoumání jevu, který reprezentuje". Jako taková je vhodná k teoretickému popisu charakteru výzkumných dat i k vztažení tohoto popisu na konkrétní situaci na Univerzitě Karlově. Metoda zakotvené teorie pracuje zejména se systematickým shromažďováním a analýzou zdrojových dat. Na základě dat jsou pak vytvářeny obecné závěry. Tato metoda v první řadě předpokládá tvorbu pojmů, jejich kategorií a vztahů mezi nimi. Sledovala jsem zejména způsob, jakým vědec s daty zacházel, jak je získával, jaký charakter (po technické i obsahové stránce) data měla a jaký byl předpokládán další způsob využití dat. Kromě této analýzy jsem provedla i analýzu struktury vzorku z hlediska zdroje dat a jejich technických vlastností, která však již přesahuje rámec tohoto článku. V následujících kapitolách jsou prezentovány výsledky analýzy metodou zakotvené teorie, a to s ohledem na kategorie definované v průběhu výzkumu. Vychází z textů zkoumaných disertačních prací.

4. Výsledky kvalitativní analýzy

Metoda zakotvené teorie předpokládá rozdělení zjištěných faktů do kategorií, v rámci, kterých jsou následně zaváděny pojmy. Vzhledem k přehledovému charakteru práce jsem nevymezovala konkrétní termíny. Pouze jsem se zaměřila na popis jednotlivých kategorií, a to zejména s důrazem na problematiku dlouhodobého ukládání digitálních dat. Kategorie jsou odvozeny na základě analýzy vzorku a na základě studia relevantní odborné literatury.

Identifikovala jsem následující oblasti:

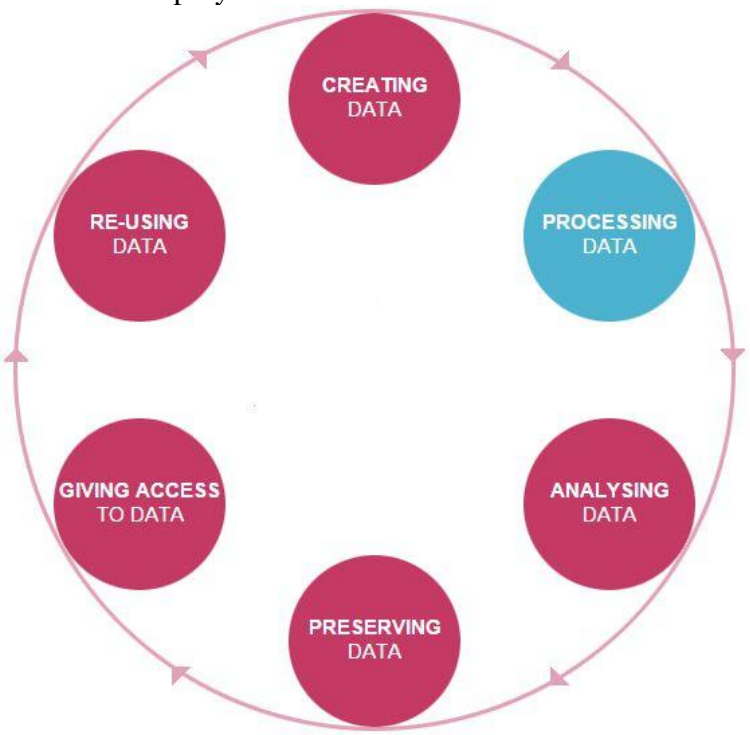
1. sběr dat a výběr výzkumného vzorku,
2. etická a právní problematika,
3. zpracování a struktura dat,
4. sdílení a archivace dat,
5. technologické otázky,
6. organizace výzkumu.

[12] STRAUSS, Anselm L a Juliet M CORBIN. *Základy kvalitativního výzkumu: postupy a techniky metody zakotvené teorie*. Boskovice: Albert, 1999. ISBN 80-85834-60-X.

Následující kapitoly přinášejí detailní popis těchto kategorií s ohledem na dlouhodobou archivaci digitálních objektů.

4.1 Sběr dat a výběr výzkumného vzorku

Fáze samotného sběru dat je pro jejich podobu a vlastnosti zcela klíčová. Z hlediska životního cyklu digitálních dat zahrnuje činnosti řazené v rámci modelu UK Data Archive (viz Obr. 2) do fází tvorby a zpracování,[13] případně do fází extrakce a transformace dle dalších zdrojů.[14] Jedná se o kategorii, která zastřešuje problematiku okolností vzniku dat a jejich převodu do digitální podoby. V metadatovém záznamu jsou informace z této kategorie zahrnuty mezi informace o původu. Zároveň také okolnosti vzniku dat slouží jako jeden z podkladů pro jejich kategorizaci z hlediska potřeb dlouhodobého uložení – tedy pro jejich rozřazení do skupin, ke kterým se vztahují různé politiky a úrovně ochrany. Informace o sběru dat a o výběru výzkumného vzorku musí být získány na základě analýzy a komunikace se skupinou tvůrců dat a s přihlédnutím k potřebám a zvyklostem určené skupiny.



Obr. 2 - model životního cyklu výzkumných dat[13]

První otázkou, které je třeba věnovat pozornost, je motivace vzniku dat, respektive motivace samotného výzkumu. Na základě obsahu disertačních prací bylo možno identifikovat tři základní motivace sběru dat:

- Sběr dat za účelem vyvození závěrů – takto je vědecký výzkum obvykle chápán. Data jsou shromažďována, aby na jejich základě mohla být potvrzena nebo vyvrácena určitá hypotéza.

[13] UK DATA ARCHIVE. Create & manage data: Research data lifecycle. In: *UK Data Archive* [online]. 2016 [cit. 2016-02-13]. Dostupné z: http://www.data-archive.ac.uk/create-manage/life-cycle

[14] LEMIRE, Daniel a Andre VELLINO. *Extracting, transforming and archiving scientific data* [online]. 2011 [cit. 2015-10-10]. Dostupné z: http://arxiv.org/abs/1108.4041v2

- Sběr dat za účelem ověření metod sběru dat – data jsou shromažďována, aby mohla být ověřena validita určitého metodického či technologického postupu nebo zařízení.
- Sběr dat je sám o sobě cílem výzkumu – data jsou jeho primárním výstupem.

Klíčovým ukazatelem je i složení a charakter výzkumného vzorku. Může se jednat o živé bytosti, přírodní jevy, společenské jevy, výsledky lidské činnosti či o objekty s jiným charakterem. Složení výzkumného vzorku ovlivňuje otázky dalšího zacházení s daty (např. otázku anonymizace). Je také klíčovým faktorem při stanovení unikátnosti výzkumu. Za unikátní jsou považována zejména data řazená do kategorie observačních. Naopak v případě experimentálních dat je do určité míry možné spoléhat na možnost opakování experimentu. Nutnost dlouhodobého uchovávání dat je však ovlivněna i náročností experimentu. Je také třeba mít na paměti, že toto rozlišení není možné aplikovat na medicínská data.

Vznik či tvorba výzkumných dat výrazně ovlivňují míru, s jakou je třeba dbát na jejich dlouhodobé uložení a obecně jeho charakter. Je zde možno rozlišit tři druhy postupů, jakými mohou být data získána.

Data mohou pocházet z externích zdrojů. Zde je zásadní věnovat pozornost míře jejich zpracování a agregace, které může být natolik významné, že výsledný dataset má charakter nové intelektuální entity. Obecně je však možno předpokládat, že externí data nemusí a vzhledem k autorsko-právním otázkám nemohou být archivována v rámci instituce, kde jsou zpracovávána. Pozornost by měla být věnována zachování vazby na externí data – například jsou-li zpracovávána data, která jsou obsahem webu (např. obsahová analýza webového sídla organizace), měla by být zachována přinejmenším vazba na jejich identifikátor (URL) a časový údaj o datu či období zpracování.

Data mohou být získána přirozenou cestou, ať už v rámci experimentu, pozorování či analýzou záznamů. Získání dat však může být různě náročné. Můžeme rozlišit náročnost:

- Technickou – zejména v otázkách nároků na přístrojové vybavení.
- Časovou – experiment či pozorování mohou probíhat dlouhodobě, či mohou být periodicky opakovány v rámci delšího časového období.
- Rozsah sběru dat – mohou být zpracovávána data získaná od velkého počtu respondentů nebo na velkém množství pokusných objektů.
- Finanční či materiální.
- Odbornou náročnost – experiment či pozorování mohou vyžadovat úzkou specializaci. V daném oboru může existovat pouze malé množství vědců schopných data získat.
- Náročnost z hlediska dostupnosti – data mohou být sbírána v odlehlé a špatně dostupné lokalitě, či k nim může být jinak omezen přístup – např. archivní záznamy.

Data mohou být získána (vytvořena) uměle jako výsledek počítačové simulace či modelu. Zde je opět třeba sledovat technickou a odbornou náročnost jejich tvorby. V tomto případě je třeba počítat s výrazným dopadem způsobu sběru dat na charakter informací, které musí být součástí metadat či dokumentace digitálního objektu určeného k dlouhodobému uchovávání.

4.2 Právní a etické otázky

Právní a etické otázky zpracování a dlouhodobého uložení výzkumných dat se z velké části odvíjejí od charakteru sběru dat, respektive od vlastností výzkumného vzorku. Klíčovou je zejména otázka, zda výzkum probíhá na lidských bytostech. Přestože výzkumy prováděné v rámci medicíny a přidružených oblastí a výzkumy z oblasti sociálních a humanitních věd jsou

z tohoto pohledu poměrně odlišné, je v každém případě třeba dbát na zachování zásad ochrany osobních údajů. Jestliže dochází k anonymizaci dat, je třeba zvážit, zda má být tato událost součástí metadatového záznamu informací o původu digitálního objektu. Součástí informací o původu mohou být i další události nebo případně i odkazy na relevantní dokumenty existující v elektronické podobě.

V případě, že vzorek dat tvoří výsledky lidské činnosti (např. umělecké předměty, či záznamy vystoupení), je třeba zachovat autorsko-právní náležitosti v oblasti duševního vlastnictví. Stejně tak je třeba věnovat pozornost i patentové problematice, která může být relevantní i v případě dat pocházejících z práce se vzorky z oblasti neživé přírody.

Zásadním důsledkem analýzy právních otázek je pak rozhodnutí o charakteru uložení a archivace – je nutné počítat s možností omezení přístupu k digitálním objektům, případě dokonce i s jejich uložením do „dark archive" tedy do archivu, který umožňuje přístup pouze pověřeným pracovníkům digitálního repozitáře.[15]

V případě, že v rámci výzkumu hrají roli etické faktory, je obvykle vyžadováno schválení etickou komisí. Status tohoto schválení nebylo na základě studovaných materiálů možno jednoznačně určit a není jisté, zda by informace o něm neměla být součástí metadatových informací o původu dat, případně jestli by opět neměla být vyžadována dokumentace tohoto rozhodnutí. Etické faktory mohou také přispívat k rozhodnutí o dlouhodobém zachování dat.

Jak již bylo výše uvedeno, velká část informací z této oblasti by měla být součástí informací o původu nebo přímo informací o právech vztahujících se k digitálnímu objektu. Jako taková by měla být součástí standardizovaného metadatového záznamu. Nicméně významná část zodpovědnosti za dodání těchto dat je na straně jejich tvůrce, tedy zodpovědné osoby v rámci výzkumného týmu. Analýza právních a etických otázek by rozhodně měla předcházet samotnému uložení objektu do repozitáře a její výstup by měl být částí plánu správy dat.

Samostatnou otázkou je pak z hlediska instituce právní vztah k externí infrastruktuře pro uložení a zpracování dat. Zde se může jednat například o subjekty, které se na zpracování dat podílejí nebo jej umožňují (např. gridová infrastruktura).

4.3 Struktura a zpracování dat

Jedním z podstatných závěrů výzkumu je skutečnost, že výzkumná data nemohou být pojímána jako zcela samostatný objekt, ale spíše jako komplexní struktura složená z více entit. Tato struktura se odvíjí od životního cyklu výzkumných dat, respektive od transformací dat, které jsou prováděny ve fázi jejich zpracování či analýzy, tak jak tyto kroky pojímá UK Data Archive[16]. Na základě výsledků výzkumu navrhuji předběžné rozlišení tří entit – dat surových, dat odvozených a dat výsledných. Je samozřejmé, že konkrétní podoba datového modelu bude závislá na dané sbírce a musí vycházet z charakteru uchovávaných dat a z potřeb dané určené skupiny. Obecně však považuji níže uvedený model za praktický základ pro další práci s výzkumnými daty i pro počáteční návrh informačních systémů pro jejich správu a ukládání.

[15] ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. 97 s.

[16] UK DATA ARCHIVE. Create & manage data: Research data lifecycle. In: *UK Data Archive* [online]. 2016 [cit. 2016-02-13]. Dostupné z: http://www.data-archive.ac.uk/create-manage/life-cycle

Data surová jsou daty výchozími. Jedná se o původní objekt, k jehož zpracování vědec přistupuje. Může být jeho tvůrcem, ale může jej také přebírat z externího zdroje. Tato data vznikají extrakcí. Lemire a Vellino[17] rozlišují v rámci fáze transformace několik druhů zpracování, z nichž většina podle mého názoru nezakládá vznik nové entity. Jedná se do určité míry o servisní zásahy, které usnadňují další zpracování dat. Příkladem může být čištění dat, ale i jejich přepis ze zvukové podoby do textové.[18] V případě slučování datových setů však již narážíme na hranici, kdy je třeba uvažovat o možném vzniku nové intelektuální entity.

Na rozdíl od typů dat, které uvádím níže, mají tato data výrazně častěji jiný než numerický (binární) formát. Na základě provedeného výzkumu lze usuzovat, že nejčetnější skupinou formátů využívaných při práci se surovými daty jsou kromě dat numerických i data obrazová.

Reálným příkladem této entity mohou být mikrofotografie biologických vzorků pořízené v rámci výzkumu. Tyto fotografie mají charakter obrazových dat (např. formát tiff). V odůvodněných případech mohou podléhat zpracování, které nezakládá vznik nové intelektuální entity – může probíhat například ořez fotografie či úprava barev.

Odvozená data jsou výsledkem výrazného intelektuálního zpracování surových dat. Je možné říci, že stejně jako data surová vznikají extrakcí, nicméně jsou extrahována z dat surových a tato vazba by měla být zachována přinejmenším v rámci metadatového záznamu. Z hlediska terminologie používané v oblasti dlouhodobého ukládání digitálních dat je možno považovat takto vzniklá data za novou intelektuální entitu. Příkladem mohou být data vzniklá automatickou či manuální analýzou mikrofotografií – výsledkem jsou data s numerickým charakterem, která často nabývají formy zpracovatelné běžným tabulkovým procesorem.

Odvozená data však mohou být nadále zpracovávána, a to zejména statistickými metodami. V této chvíli již probíhá analýza a interpretace a data jsou upravena do podoby, která se může stát přímým podkladem nebo dokonce součástí vědecké publikace. Pro tato data navrhuji termín výsledná. Variantou je i termín data statistická, neboť převážná většina těchto dat je výsledkem statistického zpracování. Tento termín by však nepostihoval případné výjimky. V kontextu výše uvedeného případu se jedná právě o výsledek netriviálního statistického zpracování.

Z hlediska zachování informací o historii vzniku digitálního objektu je třeba zvážit i možnost zachování vazby na surová data, která nenabývají digitální podoby. Může se jednat například o vazbu na papírové dokumenty s nákresy archeologických lokalit či na ručně psané zápisky. Podoba zaznamenání tohoto vztahu ovšem musí vycházet z potřeb určené skupiny. Jako naprosto dostačující se například může ukázat její zaznamenání v rámci metodologické části textu publikace.

Klíčovou otázkou, kterou si pokládá i Borgmanová,[19] je však to, kdy vlastně dataset ve smyslu nové intelektuální entity vzniká. Není možné stanovit obecnou odpověď na tuto otázku. Osoba zodpovědná za uložení výzkumných dat musí pro rozlišení použít své vlastní profesní zkušenosti

[17] LEMIRE, Daniel a Andre VELLINO. *Extracting, transforming and archiving scientific data* [online]. 2011 [cit. 2015-10-10]. Dostupné z: http://arxiv.org/abs/1108.4041v2

[18] V tomto případě vniká nový digitální objekt z pohledu dlouhodobého ukládání dat, nicméně jedná se o entitu reprezentace, která je stále součástí jedné intelektuální entity.

[19] BORGMAN, Christine L. *Big data, little data, no data: scholarship in the networked world*. Cambridge: The MIT Press, 2015. xxv, 383 stran. ISBN 978-0-262-02856-1.

stejně jako zkušenosti a požadavky, které na data klade určená skupina. Je však třeba, aby si digitální kurátor byl přinejmenším vědom toho, že výše uvedená potřeba zde je.

4.4 Sdílení dat a jejich archivace

Umožnění případného dalšího využití výzkumných dat je jednou z primárních motivací dlouhodobého ukládání a archivace výzkumných dat. Optimální archivace dat je podkladem pro jejich další využití a sdílení, a to v jakémkoli časovém i prostorovém rozsahu. Je tedy zřejmé, že požadavky na sdílená data výrazně ovlivní i charakter jejich uložení. Metody a způsoby sdílení dat v rámci určené skupiny jsou cenným zdrojem informací, ze kterých může digitální kurátor v rámci stanovení koncepce dlouhodobého ukládání vycházet.

Sdílení výzkumných dat může probíhat v rámci formální infrastruktury – tedy v rámci repozitáře či obdobného informačního systému, který je určen přímo k uložení a případnému získání výzkumných dat. Formální infrastruktura může propojovat výzkumníky na institucionální[20], národní nebo mezinárodní úrovni. V případě sdílení dat s externími subjekty je třeba dbát na dodržení nejen zákonných požadavků, ale i případných pokynů a požadavků, které na vstupní informační balíček klade daný repozitář či jiný informační systém. Podobné pokyny, jsou-li dostupné, mohou být také cenným zdrojem inspirace při tvorbě archivního informačního balíčku.

Využití externí formální infrastruktury pro sdílení dat může také výrazným způsobem ovlivnit konkrétní strategii dlouhodobého ukládání v instituci. Může být například upuštěno od dlouhodobé archivace dat, která jsou uložena v dostatečně důvěryhodném digitálním repozitáři.

Neformální sdílení probíhá zejména na základě osobních vazeb případně na základě osobní iniciativy konkrétního výzkumníka či výzkumného týmu. V některých případech může být náročné rozlišit formální a neformální infrastrukturu a je také pravděpodobné, že mnoho původně neformálních způsobů sdílení výzkumných dat se postupně transformovalo do podoby formální infrastruktury – tedy například specializovaného repozitáře.

Předpokladem neformální formy sdílení výzkumných dat je mimo jiné i jejich autoarchivace. Tato oblast by měla být při plánování strategie dlouhodobého ukládání zohledněna, a to zejména ve formě úzké spolupráce odborníků na digitální archivaci a výzkumníků samotných. Může se jednat přímo o spolupráci, ale i obecně o zvyšování informační gramotnosti v oblasti dlouhodobého ukládání.

Významným faktorem ovlivňujícím charakter sdílených a potažmo i ukládaných výzkumných dat je samotná motivace k jejich sdílení. Je možno rozlišit čtyři hlavní motivační faktory:

1) Motivace publikační politikou dané instituce potažmo státu.
2) Sdílení v rámci výzkumného týmu. Motivací je samotný charakter výzkumu.
3) Motivace vzdělávací – data jsou sdílena z populárně-naučných důvodů.
4) Sdílení v rámci vazeb a vztahů ve vědecké skupině, a to jak formálních tak neformálních. Motivací je zde obvykle možnost dalšího využití výzkumných dat.

[20] Institucionální repozitář je možno považovat za nástroj určený ke sdílení dat, a to bez ohledu na to, zda disponuje funkcemi umožňujícími dlouhodobé ukládání dat.

Za zajímavé považuji vztažení těchto motivačních faktorů k důvodům pro sdílení dat, které uvádí Borgmanová.[21] Specifikuje následující faktory:

1) Reprodukovatelnost a verifikace výzkumu.
2) Volný přístup k výsledkům výzkumu financovaného z veřejných zdrojů.
3) Možnost opakovaného výzkumu dat – nové otázky nad starými daty.
4) Obecně podpora výzkumu a vývoje.

Až na jedinou výjimku se jedná o obdobnou skupinu faktorů. V mém výčtu chybí motivace daná potřebou reprodukovatelnosti a verifikace výzkumu. Tato potřeba nebyla v žádné ze zkoumaných prací explicitně vyjádřena, jsem však přesvědčena, že implicitně je tato funkce výzkumných dat při jejich sdílení předpokládána. V určitých případech se však může jednat i o motivaci negativní. Další rozdíly jsou dány spíš odlišným cílem této analýzy – snažila jsem se najít bezprostřední faktor, který vede vědce či vědecký tým k rozhodnutí o sdílení či archivaci výzkumných dat.

Pro hlubší porozumění této oblasti je nutná analýza určené skupiny, a to zejména z hlediska způsobů jakými jsou sdílená data využívána (procesní i technologická stránka), jaké jsou jejich požadované vlastnosti a jaká jsou pravidla pro jejich popis.

4.5 Technologické otázky

Oblast čistě technologických otázek týkajících se výzkumných dat je možno rozložit do tří podskupin. V první řadě jsou tu otázky vážící se k výzkumným datům jako k digitálnímu objektu, druhou skupinou jsou otázky uložení (zejména dočasného) výzkumných dat. Třetí skupina se týká otázek vzniku a technického zpracování výzkumných dat.

Z hlediska požadavků, které na repozitář klade norma ČSN ISO 14721[22], je třeba, aby v rámci informačního objektu respektive digitálního objektu byla přítomna vysvětlující informace, tedy informace umožňující jeho interpretaci a další využití. Nedílnou součástí těchto informací jsou informace o technických parametrech digitálního objektu (v praxi obvykle nazývané technická metadata). Jejich konkrétní podobu, obsah a metadatový profil je možno určit až na základě analýzy konkrétních digitálních objektů a požadavků určené skupiny. Obecně je však možno říci, že bude třeba sledovat zejména formát, ve kterém jsou daná výzkumná data ukládána, a jejich velikost. Zejména (ale ne pouze) v případě komputačních dat je třeba věnovat pozornost i použitému programovacímu jazyku a případně i operačnímu systému. Jedním z prvních kroků při dlouhodobém ukládání výzkumných dat by měla být analýza a určení signifikantních vlastností digitálních objektů, a to jak na základě technického charakteru, tak i na základě kontextu v jakém data vznikají a jsou využívána.[23]

---

[21] BORGMAN, Christine L. The conundrum of sharing research data. *Journal of the American society for information science and technology* [online]. 2012, **63**(6), s. 1059-1078 [cit. 2016-02-13]. DOI: 10.1002/asi.22634. ISSN 15322882. Dostupné z: http://doi.wiley.com/10.1002/asi.22634

[22] ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. 97 s.

[23] DAPPERT, Angela a Adam FARQUHAR. Significance is in the eye of the stakeholder. In: *Research and advanced technology for digital libraries: 13th European conference*, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. [online]. 2009. s. 297 [cit. 2015-01-17]. DOI: 10.1007/978-3-642-04346-8_29. Dostupné z: http://link.springer.com/10.1007/978-3-642-04346-8_29

Informace o vzniku a o případných úpravách výzkumných dat by měly být součástí informace o původu. Opět i zde platí, že přesná podoba může být určena až na základě prozkoumání konkrétních digitálních objektů a určené skupiny. Nicméně klíčové zde budou zejména informace o použitém snímacím či nahrávacím zařízení, o jeho nastavení a o softwaru použitém jak pro samotné pořízení dat, tak pro další zpracování. Vhodné je zaznamenat i informace o konkrétních krocích zpracování.

Ve většině případů je možné předpokládat, že výzkumná data jsou po dobu výzkumu samotného (a často i po jeho skončení) uložena v nějaké formě dočasného úložiště. Může se jednat o pevný disk osobního počítače, ale i o cloudové úložiště nebo o externí záznamové zařízení. Striktně vzato problematika dočasných úložišť není součástí dlouhodobého ukládání digitálních objektů. Zejména na institucionální úrovni však nemůže být zcela ignorována, protože má vliv na podobu vstupních balíčků ukládaných do repozitáře. Politika zabezpečení a provozu dočasných úložišť by měla být vytvářena s ohledem na politiku dlouhodobého ukládání výzkumných dat.

4.6 Organizace výzkumu

Jako poslední skupinu otázek jsem identifikovala problematiku samotné organizace výzkumu. Ve srovnání s výše uvedenými se jedná spíše o méně významný okruh otázek, který se dotýká zejména praktické organizace vstupu dat do systému.

V rámci příjmu výzkumných dat do repozitáře – tedy v rámci tvorby vstupního balíčku – je třeba komunikovat s tvůrci informace. V případě výzkumných dat je třeba zjistit složení výzkumného týmu a identifikovat jak kontaktní osobu, tak osobu, která bude schopna zodpovědět případné technické otázky. Je třeba zvážit, zda není vhodné uchovat kontakt na tyto osoby v rámci metadatového záznamu archivního balíčku. V případě komplikovanějších výzkumných projektů považuji za užitečné provést i analýzu interní komunikace výzkumných dat v rámci vědeckého týmu.

Za účelnou považuji i analýzu obecně platných zvyklostí v daném oboru. Může se jednat o otázky týkající se užívaného jazyka či terminologie a obecně o informace, které se mohou stát součástí vysvětlující informace.

5. Souhrnné závěry

Z textů zkoumaných disertačních prací je patrná potřeba sdílet, a tím pádem i ukládat získaná data, stejně jako skutečnost, že k využití již existujících dat běžně dochází. Jedná se o aktuální oblast zájmu, která však ještě není řádně zakotvena ve struktuře instituce. Není jasné, zda jsou za uložení a sdílení výzkumných dat zodpovědné knihovny, archivy či zda se jedná o specifickou složku vyžadující samostatnou infrastrukturu. Je pravděpodobné, že v budoucnu se na archivaci výzkumných dat budou podílet všechny tyto složky, a proto by mělo být v jejich zájmu věnovat se hlubší analýze této problematiky.

Analýza ukázala, že dělení výzkumných dat dle jejich zdroje je do určité míry relevantní pro archivaci digitálních objektů. Výjimečný charakter mají zejména komputační data. Nicméně toto dělení není možné použít jako jednoznačný podklad pro přiřazení konkrétní politiky ochrany. Přesněji řečeno, vhodnějším kandidátem by byl způsob sběru dat, který je možno vymezit jemněji a odlišit tak výrazně různé skupiny dat.

Klíčový význam má také životní cyklus digitálních výzkumných dat, respektive vztah mezi daty surovými a odvozenými. Ve většině případů není možno pojímat výzkumná data jako samostatnou a jednoduchou entitu. Popis jejich struktury je třeba založit na vztazích, které vznikají na základě zpracování či transformace surových dat.

Kvalitativní analýza vzorku potvrdila klíčovou roli analýzy určené skupiny ve smyslu, v jakém ji chápe norma ČSN ISO 14721[24]. Nabízí se několik možností zkoumání skupiny. První z nich je přímá analýza, která může využívat klasických metod používaných v sociálních vědách – tedy zejména dotazníkového šetření a rozhovorů. Varianta použitá v rámci mého výzkumu – tedy obsahová analýza vědeckých výstupů má svá omezení a je použitelná spíše jako výchozí bod pro další zkoumání. Přímou analýzu je možno doplnit průzkumem mechanismů sdílení výzkumných dat v rámci určené skupiny. Významná může být i spolupráce s výzkumnými týmy, a to zejména při budování dočasných či pracovních úložišť digitálních objektů.

Příjmu dat do repozitáře by měla předcházet jak analýza určené skupiny, tak analýza struktury výzkumných dat s přihlédnutím k jejich životnímu cyklu v rámci výzkumu. Pouze na základě relevantních informací o určené skupině a o struktuře výzkumných dat je možno vytvořit vhodný datový model struktury digitálního objektu a optimální metadatový profil pro uložení relevantních informací.

[24] ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. 97 s.

Použitá literatura

BORGMAN, Christine L. The conundrum of sharing research data. *Journal of the American society for information science and technology* [online]. 2012, **63**(6), s. 1059-1078 [cit. 2016-02-13]. DOI: 10.1002/asi.22634. ISSN 15322882. Dostupné z: http://doi.wiley.com/10.1002/asi.22634

BORGMAN, Christine L. *Big data, little data, no data: scholarship in the networked world*. Cambridge: The MIT Press, 2015. xxv, 383 stran. ISBN 978-0-262-02856-1.

CAPLAN, Priscilla. Chapter 1: What is digital preservation? *Library technology reports* [online]. 2008, **44**(2) [cit. 2016-03-18]. ISSN 0024-2586. [cit. 2014-12-27]. DOI: 10.5860/ltr.44n2. Dostupné z: https://journals.ala.org/ltr/article/view/4224/4809

CAPLAN, Priscilla. *Understanding PREMIS* [online]. Washington: Library of Congres, 2009 [cit. 2016-02-13]. Dostupné z: http://www.loc.gov/standards/premis/understanding-premis.pdf

ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. 97 s.

DAPPERT, Angela a Adam FARQUHAR. Significance is in the eye of the stakeholder. In: *Research and advanced technology for digital libraries: 13th European conference*, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. [online]. 2009. s. 297 [cit. 2015-01-17]. DOI: 10.1007/978-3-642-04346-8_29. Dostupné z: http://link.springer.com/10.1007/978-3-642-04346-8_29

ECONOMIC AND SOCIAL RESEARCH COUNCIL. Research data policy. In: *Economic and social research council* [online]. ESRC [online]. 2015 [cit. 2016-01-27]. Dostupné z: http://www.esrc.ac.uk/funding/guidance-for-grant-holders/research-data-policy/

HENDL, Jan. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Vyd. 2., opr. Praha: Portál, 2006. 583 s. ISBN 80-7367-123-9.

LEMIRE, Daniel a Andre VELLINO. *Extracting, transforming and archiving scientific data* [online]. 2011 [cit. 2015-10-10]. Dostupné z: http://arxiv.org/abs/1108.4041v2

NATIONAL SCIENCE BOARD. *Long-lived digital data collections: Enabling research and education in the 21st century* [online]. 2005. Arlington: National Science Foundation [cit. 2016-02-13]. Dostupné z: http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

PAVLÁSKOVÁ, Eliška. *Výzkumná data na Univerzitě Karlově v Praze*. 16. konference Archivy, knihovny, muzea v digitálním světě 2015.

STRAUSS, Anselm L a Juliet M CORBIN. *Základy kvalitativního výzkumu: postupy a techniky metody zakotvené teorie*. Boskovice: Albert, 1999. ISBN 80-85834-60-X.

UK DATA ARCHIVE. Create & manage data: Research data lifecycle. In: *UK Data Archive* [online]. 2016 [cit. 2016-02-13]. Dostupné z: http://www.data-archive.ac.uk/create-manage/life-cycle