

Journal of eScience Librarianship

eScience in Action

Data Curation in Practice: Extract Tabular Data from PDF Files Using a Data Analytics Tool

Allis J. Choi and Xuying Xin

The Pennsylvania State University, University Park, PA, USA

Abstract

Data curation is the process of managing data to make it available for reuse and preservation and to allow FAIR (findable, accessible, interoperable, reusable) uses. It is an important part of the research lifecycle as researchers are often either required by funders or generally encouraged to preserve the dataset and make it discoverable and reusable. This has been especially important as the Open Access (OA) policy is being implemented in many institutions across the nation. In facilitating research data discovery and enhancing its easier reuse, an efficient data repository and its data curation play key roles. In this article, we briefly discuss the local institutional repository at Penn State University and the general data curation practices we adopt for the deposited files and datasets, then we focus on a data analytics tool that has recently been applied to extract tabular data from PDF files. This is an enhancement to the existing data curation practices as it adds additional tabular data to deposits with PDF files where tables are often embedded and not easily reused.

Correspondence: Xuying Xin: xzx1@psu.edu

Received: April 8, 2021 Accepted: June 4, 2021 Published: August 11, 2021

Data Availability: The datasets analyzed during the current study are available at https://doi.org/10.7554/elife.44898.

Disclosures: The authors report no conflict of interest.

Copyright: © 2021 Choi & Xin. This is an open access article licensed under the terms of the Creative Commons Attribution License.

Introduction

Launched in 2012, our institutional repository, ScholarSphere,¹ enables University faculty, students and staff to deposit and actively manage their scholarly works, and share them with the university community and the world.² The local data curation team adopts general data curation practices, CURATE(D) steps (Check, Understand, Request, Augment, Transform, Evaluate, Document) to curate files and datasets in various research areas (STEM, Liberal Arts, Social Sciences) and in various formats including tabular data, image data, software code, etc. We partner with Data Curation Network (DCN), a network of over ten institutions with shared data curation expertise while also providing normalized data curation practices and professional development training (Johnston et al. 2018). Our partnership with DCN enhances data curation expertise with additional support when there is a lack of such expertise or there is a peak time demand in the local data curation work. With the increasing number of deposits to the newer version ScholarSphere that was launched recently, our team has created a workflow for managing data curation process and it has greatly helped local curators collaborate and curate data efficiently. Data Science and Technologies advanced tremendously in the past few years, and brought innovations to almost every segment, including retail, finance, government, education, information technology, and research libraries. Recently, our libraries have started a data analytics and visualization service to support research community of faculty, staff and students with data analysis and visualization for their research projects. One of the top data analytics tools we have been using for visualization, Microsoft Power BI Desktop, has found its additional usage in data curation. There are two versions of Power BI: Power Desktop and Power BI online. Power BI Desktop (https://www.microsoft.com/enus/download/details.aspx?id=58494) which is free to anyone, is the one that we use for extracting tables from PDF files. This desktop version of the application which runs only on Windows computers, provides over sixty data connectors including one for PDF files. Power BI Online does not support the data extraction feature but it does support data visualization and data sharing via its online server. The reason we chose to use Power BI Desktop for data extraction from PDF files is that it is easy to use, and it is free to anyone; Power BI Online is not free to everyone but free to users within organizations that have purchased a group Office 365 license which many universities have. PDF files are the majority type of files $(\sim 80\%)$ deposited to our institutional repository. Extracting the embedded tables and saving them as CSV files or Excel worksheets increases the reusability of the research data that otherwise is embedded and not readily available for reuse.

Extracting Computational Data from PDF files

The following graph demonstrates the process and steps (Figure 1) for extracting tabular data from PDF files by using Power BI Desktop, a popular analytics tool for data analysis and visualization.

¹ https://scholarsphere.psu.edu

² https://news.psu.edu/story/146447/2012/09/20/scholarsphere-repository-promotes-research-sharing-and-discovery



Figure 1: The process and tool for extracting tabular data from PDF files deposited to a local repository.

The process starts with finding the deposits with PDF files such as research papers, articles and reports in the local repository (Figure 2) by using the "Work Type" search, then downloading and saving the PDF files to a local folder for the software tool to access for data extraction.

📃 🛛 Welcome to ScholarSphere	LIMIT YOUR SEA	ARCH		V1 PUBLISHED -	Edit V1			
	Work Type		Computational 3D	D Histological				
SCHOLARSPHERE UPDATES	Dataset	1,526	Phenotyping of Wh	sh by X-				
We have released a new version of ScholarSphere! The new platform includes an improved interface for uploading your work and better support for large files. We've also added support for 'draft' works and work versions. Additional features will be released in the coming months. read more	Article	1,108	ray Histotomograp	-				
	Book	1,013	Abstract: Organismal phenotypes freq	rgan systems.				
	Poster	551	Histology is a powerful way to detect or descriptive and subjective. To determine	cellular and tissue phenotypes, but is largely ine how synchrotron based X-ray micro-				
	Map Or Car-	518	tomography (micro-CT) can yield 3-dimensional whole-organism images sui for quantitative histological phenotyping, we scanned whole zebrafish, a sm wertabete medel with diverse tirsues at -1 micron unvel excelutions. Using					
	Material		verteorate model with overse ussues, at ~1-micron voxel resolutions. Using micro-CT optimized for cellular characterization (histotomography), brain nuclei were computationally segmented and assigned to brain regions. Shape and					
BROWSE AND SEARCH FOR WORKS	Research	473	volume were computed for populations of nuclei such as those of motor neurons and red blood cells. Striking individual phenotypic variation was apparent from					
	A Paper	Paper color maps of computed cell density detection of phenotypes that requir			nography allows the n multiple planes.			
Search for titles, keywords, resource types, creator	Presentation	457	We expect the computational and visu provided by histotomography to be us	al insights into 3D tissue eful for reference atlases	architecture s, hypothesis			
	Report	432	generation, comprehensive organisma	l screens, and diagnostic	:5.			
OR	Unspecified	351	FILES					
	Other	343						
Browse & Filter All Works	more »		Zebrafish_histotomography_eLife.pdf 10.3 MB application/pdf 2021-04-06					

Figure 2: The interface of our local institutional repository, ScholarSphere, and an example PDF file with tables to be extracted.

The next step is to launch the MS Power BI Desktop (https://www.microsoft.com/ en-us/download/details.aspx?id=58494) that can be downloaded on a Windows computer for free for anyone. The application has a data connector for PDF files. Click "Get data," Select "PDF," and locate the file in the folder. The tool detects and lists all the tables it identified; it also allows to preview each table to be extracted. In Figure 3, a screenshot of an embedded table in the PDF file (Ding et al. 2019) is placed next to the preview version in Power BI Desktop for comparison purpose. The information in both tables is consistent. Multiple tables can be extracted at once after quality checks, some renaming might be needed for the extracted tables. One tip for locating tables in PDF files to compare with the preview versions in Power BI Desktop is to use 'Ctrl' + 'F' and the keyword 'table' to track down all the tables instead of examining the whole article manually by mouse scrolling.

Home Insert	Modeling View Hel	p		Navigator					C	
Cut Copy Format painter	Excel Power BI SQL Enter datasets Server data s	Recent ources data v	New C measure me	Display Options *	م ۵	Table010 (Page 20) Preview downloaded on Materbrainfunctions Colume3 Brainragtonyolumes(mer19)				
board	Data	Queries Insert	Calculatio	Table009 (Page 20)	^	Specimennumber	1	2	3	
			~ 1	✓ 1 Table010 (Page 20)		OlfactoryEpithelium	293251	171459	164338	
	Cot Data		^	Table011 (Page 23)		Telencephalon	1117500	1021518	956493	
	Get Data			Table012 (Page 25)		Diencephalon	1440447	1300518	1097698	
				Table013 (Page 25)		Hypothalamus	813210	702116	732073	
	Search	All		Para Cont		Mesencephalon	4353108	3442415	3663909	
	All	Excel	~			Metencephalon	4082069	4636451	365409	
	201	P Int/CSV				WhiteMatter	4783477	4358188	4119302	
	File	E lext/CSV		E E Pageoos		SpinalCord	78092	97609	117260	
	Database	🖻 XML		Page004		Total	18185716	16003973	15217359	
	Power Platform			D Base006						
		Folder								
	Azure		- L L							
	Online Services	PDF		Table 3. Brain Volumes by Regio	on for t	ive dpt Zebratist	n Brain.			
	Other	SharePoint folder		Major divisions of the zebrafish t	orain ai	nd their volumes	s (μm²) are s	nown acro	iss five zebrafish. Mea	
	ound	SOI Server database		ative standard deviation ± 11.9%	h					
		Set Server Galabase		Major brain divisions				Brain	region volumes (µm ³)	
		Access database		Specimen number	1		2		3	
		SQL Server Analysis Services database		Olfactory Epithelium	293,2	251	171,459		164,338	
		Oracle database		Telencephalon	1,117,	500	1,021,518		956,493	
		IBM Db2 database		Diencephalon	1,440,	447	1,300,518		1,097,698	
		IBM Informix database (Beta)		Hypothalamus	813,2	210	702,116		732,073	
		BM Netezza		Mesencephalon	4,353,	108	3,442,415		3,663,909	
		MASOI database		Metencephalon	324,5	562	273,699		388,409	
		injuge outdouse	~	Myelencephalon	4,982,	069	4,636,451		3,977,877	
		PostgreSQL database		White Matter	4,783,	477	4,358,188		4,119,302	
				Spinal Cord	78,0	92	97,609		117,260	
	Certified Connectors	Connect Car	ncel	Total	18,185	,716	16,003,973		15,217,359	
I				DOI: https://doi.org/10.7554/eLife.44	4898.027	1				

Figure 2: Demonstration for importing a PDF file into Power BI Desktop and extracting tabular data embedded in the PDF file.

The last step is to extract the tables by clicking "Load" in the data transformation dialog window, then the tables will show up in the "Data" section with a table icon on the right side of the application window in Power BI Desktop; right click the table to use the "Copy Table" option to paste and save the tables in Excel as CSV format. We also reference the extracted tables back to the work page in the repository that contains the related PDF files. Quality check is very important for the data extraction process when using this software tool since some math equations in the embedded tables of PDF files can complicate the process and cause inaccurate data extraction.

Conclusion

We have discussed a data analytics and visualization tool for extracting tabular data from PDF files that are deposited to the local institutional repository. These files take about 80% deposits that the scholars have made for sharing their research products. When tabular data is embedded in PDF files, it is not easy to reuse. With this free software tool, data can be easily extracted in batch. This enhances the data curation work we do by increasing the amount of data available

to be reused. Another data analytics tool, Tableau, also has the feature for extracting tabular data from PDF files; however, it is not always free. The recently launched newer version institutional repository, ScholarSphere, provides many new features that improve the efficiency of the data curation work process. We have seen an increasing number of researchers choosing to share their research products via this institutional repository.

Acknowledgements

We would like to thank the data curation team at the Penn State University Libraries for the discussions and support for this work, the Data Curation Network (DCN) for all the training and shared expertise in research data curation, Ally Laird, Paulina Krys, and Tara Anthony from the Research Informatics and Publishing at the Penn State University Libraries for feedback on the manuscript, and Dr. Keith C. Cheng from the Penn State College of Medicine for allowing us to use his research article to demonstrate the data extraction process with the data analytics tool.

Data Availability

The datasets analyzed during the current study are available at https://doi.org/10.7554/elife.44898.

References

Ding, Yifu, Daniel J Vanselow, Maksim A Yakovlev, Spencer R Katz, Alex Y Lin, Darin P Clark, Phillip Vargas, et al. 2019. "Computational 3D Histological Phenotyping of Whole Zebrafish by X-Ray Histotomography." *ELife* 8(May). https://doi.org/10.7554/elife.44898

Johnston, Lisa R, Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, Claire Stewart, et al. 2018. "Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data." *International Journal of Digital Curation* 13(1): 125–140. https://doi.org/10.2218/ijdc.v13i1.616