Next-Generation PDF Files

Marrying Data Sets to Academic Publications

Overview

Although most applications use PDF only for supplemental content, such as user manuals or digital publications, *interactive* PDF technology would significantly expand the use-cases where PDF documents can be an essential part of the User Interface for software applications. The relatively restricted adoption of PDF is driven partly because traditional PDF offers only limited options for interactivity, through features such as PDF scripting, and in fact scripting itself is limited because many PDF users/viewers disable scripting for security reasons. Another reason for the restricted use of PDF interactivity is that PDF generation tools are not programmed to identify document locations and actions where interactive capabilities are needed. Such limitations can, however, be rectified by embedding metadata within PDF documents which identify special actions to associate with individual PDF coordinate regions, and by implementing callback handlers in PDF viewers to support user actions and context menus linked to those coordinate regions.

PDF files augmented with these extra features can make the PDF experience much more interactive and user-friendly, allowing PDFs to play a larger role in application development. For exampe, PDF documents may be connected to external applications that provide graphics and multimedia, or application-specific features such as diagram/plotting engines and practice test questions.

More generally, interactive PDF technology can integrate PDF viewers with a wide range of software applications. This interoperability can be achieved by encoding and sharing the textual content from which a PDF document is generated, so that this raw text is available to user-action handlers within the PDF application, and by annotating external data and code which may be linked to the PDF application so as to cross-reference these external resources with the document text. The corresponding relationship between text encoding and data/code annotation is outlined below.

Technology Stack

1. **A New Text Encoding Protocol** — Introducing a novel text encoding for document markup, including PDF generation, which produces user-friendly and machine-readable text representation and metadata that may be embedded as a PDF attachment. The resulting supplemental material, which a compatible PDF viewer can automatically extract, would consist of two parts: (1) a direct encoding of document text in the form of a character stream with a fixed character size (e.g. one byte per character); and (2) metadata representing the character positions/ranges and PDF coordinates of document content such as sentences, paragraphs, citations, keywords, and figures/graphics.

   With respect to document text, having a character stream (or several character stream "layers") as a direct representation of textual content — rather than "scraping" characters from the PDF directly — leads to more accurate searching and UI features (e.g. copy/select). For example, with one single context menu action users can copy an entire sentence/paragraph with just one click. This is made possible because the character streams employ a customizable encoding (instead of a generic format like Unicode) so that the document can specify ahead of time how text segments should be translated into character sequences for searches or selection-copy features; notably, how

to handle foreign letters, equations, footnotes, and so forth, details which ordinarily complicate basic functionality such as "copy a range of text to the clipboard." Fixed-width character encoding ensures that all text segments can be identified by simple numeric ranges, and text can be decoded by iterating over one character at a time in isolation, in contrast to Unicode multi-bytes or XML character entities. With respect to metadata, supplemental files encode annotations which map character-stream positions to PDF coordinates for document content that would be a target for user interactions, such as sentence boundaries (to enable automatic sentence-copy features), figures/graphics (for context menus providing figure information), citations (for capabilities such as searching cited publications' DOI) and so forth. We have put together a demo PDF viewer which reads the character and metadata files and uses this data to provide new capabilities that are not found in other PDF programs.

2. **Dataset Integration and "Microcitations"** — When integrating data sets with PDF documents, our text-encoding protocol is able to support fine-grained cross-references between individual parts of the data set and individual locations or regions of the document. This contrasts with the conventional approach wherein connections between data sets and publications are noted only coarsely, such as when a "Supplemental Materials" section of a paper provides a link to download an accompanying data set, but only with a cursory overview of what is inside the data set. By contrast, our Scientific Data Repository Framework (**SDRF**) allows granular cross-references where, for example, a concept embodied in a data set (for instance, a statistical parameter instantiated by a table column or datatype field) can be mapped to the sentence in the publication where that concept is defined or is first introduced. Likewise, tables or figures diagrammed in a document could be linked to the relevant part of their accompanying data set, such as individual tables which generate a statistical plot, or source code files where statistical distributions are calculated. So, in order to support these granular cross-references — which are a form of "microcitation" as this term is used in science — we have prototyped novel "dataset creator" tools that build customized "dataset applications" for viewing and analyzing dataset contents. In the spirit of initiatives such as Research Objects and FAIRsharing (Findable, Accessible, Interoperable, Reusable) — standards which have been prioritized by institutions such as the NIH, the Bill and Melinda Gates Foundation, and the Chan Zuckerberg Initiative — dataset applications combine code, data, and visualization/GUI tools and, moreover, can be launched directly from the PDF viewer itself. A standardized protocol cross-referencing PDF files with data sets (annotating specific parts of the document and the data set which are conceptually related) could also include specifications for launching external applications to view embedded multi-media content, such as 3D graphics or video files. In effect, support for dataset annotations could be merged with an annotation system whereby files embedded in a PDF document — potentially linked to visible PDF elements such as a 2D figure representing a view onto a 3D model or a still frame from a video — are annotated with metadata allowing the PDF viewer to export resources to an external application.

3. **Cloud Hosting** — Projecting forward, with the idea that authors and publishers start to provide granularly connected documents and data sets, it will be necessary to implement new hosting platforms which can leverage such document/data integration protocols. At present, data-hosting sites (such as Dryad, Sciverse, or Open Science) and document repositories (such as Science Direct, Springer Nature, or ScienceOpen) usually operate in parallel, but a next generation of content providers may need to host both data and documents simultaneously, in an integrated fashion, allowing for searches across both publications and data sets. This means that data sets have to be indexed and searchable as well. When a document is linked to a data set, the document viewer (canonically a PDF viewer) would then have the option of accessing the document source and retrieving dataset information — for example, identifying software prerequisites for using the data set and explaining to readers how they can acquire and access the dataset contents — and once the data set is downloaded and the relevant software is identified (either as an application embedded in the data set or as an external program) the PDF viewer could then connect with that software component as part of its user-interaction interface. These kinds of workflows would require APIs and multi-application networking protocols that could be concretized by providing

cloud-hosting services or projects such as a reusable docker container, which demonstrate API and protocol implementations as well as document and dataset hosting capabilities.

A New Protocol for Sharing Scientific Data

Linguistic Technology Systems (LTS) has developing a novel Dataset Creator (dsC) to encompass data models, publishing guidelines, and code libraries for deploying open-access research data sets associated with scientific publications. Nowadays there are many general-purpose and domain-specific portals hosting scientific data; there are also several available formats for describing and encoding scientific data, such as Research Objects, **schema.org/Dataset**, Digital Curation Center (DCC), SciData, BioCoder, and MIBBI (Minimum Information for Biological and Biomedical Investigations). The purpose of dsC is to merge these different data-set formats into a unified, overarching standard which can be adapted to different publishing houses and pipelines.

In order to conform to current specifications — such as FAIRsharing, the Bill and Melinda Gates Foundation guidelines for authors (https://gatesopenresearch.org/for-authors/data-guidelines), or the Chan Zuckerberg "Invest in Open Infratructure" Initiative — many publications need to be accompanied by well-documented open-access data sets which include metadata and supporting files to help researchers properly access, visualize, and reuse the data, and to provide software which has the correct features to load and display the relevant raw data files. In response to these specifications, LTS provides a "Scientific Data Repostory Framework" (SDRF) so that data sets constructed via dsC can embody the protocols recommended by FAIRsharing and the Gates or Zuckerberg initiatives.

Current data-publishing standards demand a more rigorous data-curation process than simply hosting open-access data sets and referencing them indirectly in published articles. The first three screenshots below demonstrate older methods for syncing articles and data sets, which obfuscate the source of research data and make it difficult for subsequent researchers to obtain and reuse this data. These screenshots (from one example of a traditional publication) exemplify limitations in the publishing workflow when articles are not closely linked to accompanying data sets.

"Problem" Example:

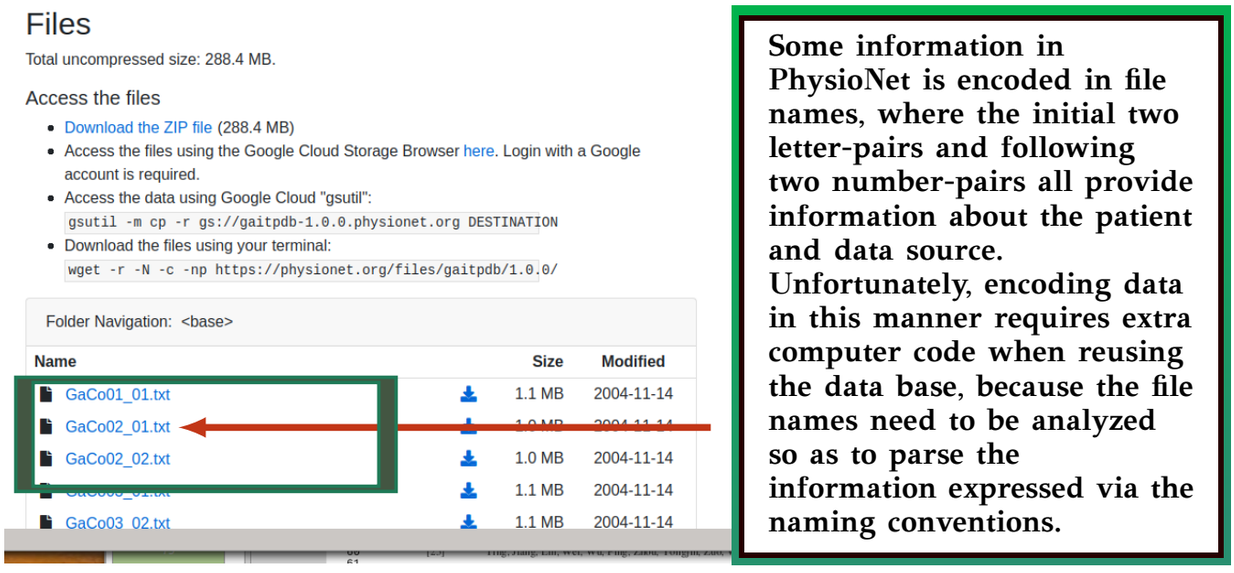

Figure 1: Extracting Information Encoded in File Names

3.2. Data preprocessing and Feature extraction

Since the data is extracted using different signal processing methods, it ranges diversely. This contributes to inadequate learning procedures. Consequently, to get started with the task, we apply rescaling or in a more common term, min-max normalization. Using this method, the data is scaled in a specified range, and here we scale the features to the [0, 1] range.

Table 1: A summary of reviewed datasets.

Data type	Description	Study
Brain MRI	In this retrospective study, we enrolled 56 patients and 28 healthy control subjects.	[9]
GAIT	This database contains measures of gait from 93 patients with idiopathic PD (mean age 66.3 years; 63% men), and 73 healthy controls (mean age: 66.3 years; 55% men).	[10]
GAIT	303 subjects were recruited from the "Incidence of Cognitive Impairment in Cohorts wit Longitudinal Evaluation-GAIT" (ICICLE-GAIT) study.	[11]
Vocal Features	UCI Parkinson's Disease Classification.	[13]
Vocal Features	The dataset range of biomedical voice measurements from 31 people, where 23 peopl are showing Parkinson's disease. -UCI Parkinson Speech Dataset with Multiple Types of Sound Recordings Data Set	[14]
Vocal Features	The database consisted of 23 columns and 197 rows. The dataset was created by Ma Little of the University of Oxford, in collaboration with the National Centre for Voice an Speech, Denver, Colorado, who recorded the speech signals. This dataset is composed of a range of biomedical voice measurements from 31 people 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure	[15], [17]

The authors provide citations for data sets used in their analyses ...

Figure 2: Table Listing Analyzed Data Sets in the Parkinson's Article

[8] Sakar, C. Okan, Gorkem Serbes, Aysegul Gunduz, Hunkar C. Tunc, Hatice Nizam, Betul Erdogdu Sakar, Melih Tütüncü, Tarkan Aydin, M. Erdem Isenkul, and Hulya Apaydin. "A Comparative Analysis of Speech Signal Processing Algorithms for Parkinson's Disease Classification and the Use of the Tunable Q-factor Wavelet Transform." *Applied Soft Computing* 74 (2019): 255-63. doi:10.1016/j.asoc.2018.10.022.

[9] Salvatore, C., A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri, M. Lopez, G. Arabia, M. Morelli, M.c. Gilardi, and A. Quattrone. "Machine Learning on Brain MRI Data for Differential Diagnosis of Parkinson's Disease and Progressive Supranuclear Palsy." *Journal of Neuroscience Methods* 222 (2014): 230-37. doi:10.1016/j.jneumeth.2013.11.016.

[10] "Gait in Parkinson's Disease." Gait in Parkinson's Disease V1.0.0. February 25, 2008. https://physionet.org/content/gaitpdb/1.0.0/

[11] Rehman, Rana Zia Ur, Silvia Del Din, Yu Guan, Alison J. Yarnall, Jian Qing Shi, and Lynn Rochester. "Selecting Clinically Relevant Gait Characteristics for Classification of Early Parkinson's Disease: A Comprehensive Machine Learning Approach." Scientific Reports 9, no. 1 (2019). doi:10.1038/s41598-019-53656-7.

[12] Goetz, Christopher G. "The History of Parkinson's Disease: Early Clinical Descriptions and Neurological Therapies." *Cold Spring Harbor Perspectives in Medicine*, Cold Spring Harbor Laboratory Press, Sept. 2011.

[13] UCI Machine Learning Repository: Parkinson's Disease Classification Data Set. https://archive.ics.uci.edu/ml/datasets/Parkinson's Disease Classification.

[14] Sriram, Tarigoppula & Rao, M. & Narayana, G & Vital, T. & Dowluru, Kalat[illegible] V GK. (2013). Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms. IJEIT. 3. 212-215.

[15] R. Das, "A Comparison of multiple classification methods for diagnosis of Parkinson disease." Expert Systems with Applications, vol. 37, no. 2, pp. 1568-1572, 2010.

[16] Erdogdu Sakar, Betul et al. "Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease." *PloS one* vol. 12,8 e0182428. 9 Aug. 2017, doi:10.1371/journal.pone.0182428

[17] M. Peker, B. Şen, D. Delen, Computer-aided diagnosis of Parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm, J. Healthcare Eng. 6 (3) (2015) 281–302

[18] Ahlrichs, Claas, and Michael Lawo. "Parkinson's Disease Motor Symptoms in Machine Learning: A Review." Health Informatics - An International Journal 2, no. 4 (2013): 1-18. doi:10.5121/hiij.2013.2401.

[19] Khoury, Nicolas, Ferhat Attal, Yacine Amirat, Abdelghani Chibani and Samer Mohammed. "CDTW-based classification for Parkinson's Disease diagnosis." ESANN (2018).

[20] Brooks, David J. "Neuroimaging in Parkinson's Disease." NeuroRX 1, no. 2 (2004): 243-54. doi:10.1602/neurorx.1.2.243.

[21] Mohammad, Roohi, and Fatima Mubarak. "Neuroimaging in Parkinson Disease." Parkinson's Disease and Beyond - A Neurocognitive Approach, 2019. doi:10.5772/intechopen.82308.

[22] A. Kazeminejad, S. Golbabaei and H. Soltanian-Zadeh, "Graph theoretical metrics and machine learning for diagnosis of Parkinson's disease using rs-fMRI," 2017 Artificial Intelligence and Signal Processing Conference (AISP), Shiraz, 2017, pp. 134-139.

[23] Shiiba T, Arimura Y, Nagano M, Takahashi T, Takaki A. "Improvement of classification performance of Parkinson's disease using shape features for machine learning on dopamine transporter single photon emission computed tomography." PLoS ONE 15(1): e0228289, (2020). doi: 10.1371/journal.pone.0228289.

[24] Xu, Jiahang, Jiao, Huang, Yechong, Luo, Xu, Qian, Li, Ling, Liu, Zuo, Wu, Ping, and Xiahai. "A Fully Automatic Framework for Parkinson's Disease Diagnosis by Multi-Modality Images." Frontiers. August 05, 2019.

[25] Ting, Jiang, Lin, Wei, Wu, Ping, Zhou, Yongjin, Zuo, Wang, Jian, Yan, Zhuangzhi, Shi, Kuangyu,

but while some data sets are directly available through the bibliography, others have to be located by reading the cited articles ...

Figure 3: Bibliography (With Data Set Hyperrefs) in the Parkinson's Article

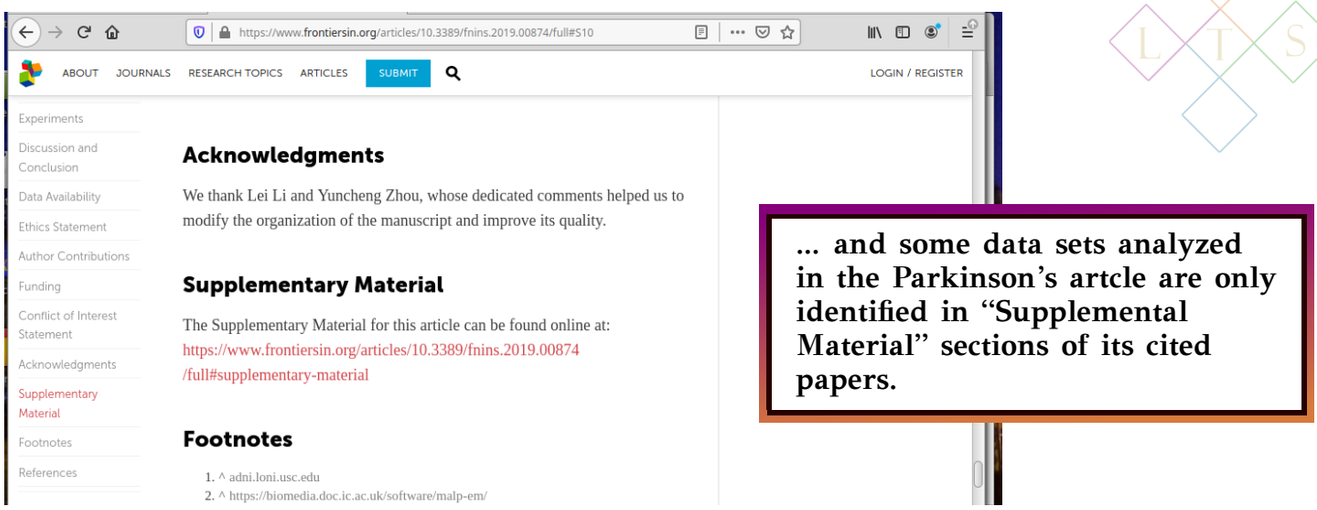

Figure 4: Indirectly Locating Data Sets from Cited Papers

"Solution" Example:

The final three screenshots below, by contrast, show an example of a data set constructed using a version of dsC and SDRF. This data set, which accompanies a recently published article in the International Journal of Speech Technology, evinces a more rigorous integration between the published article and the various resources contained within the data set. The data set includes a custom application providing access to its raw data files, and also includes several PDF files which explain and analyze the raw data in more detail than is possible within the scope of the parent article. Both the data-set application and these supplemental files refer back to the parent article, providing multiple pathways where the parent article may be found and cited (potentially boosting downloads and impact factor for the publication).

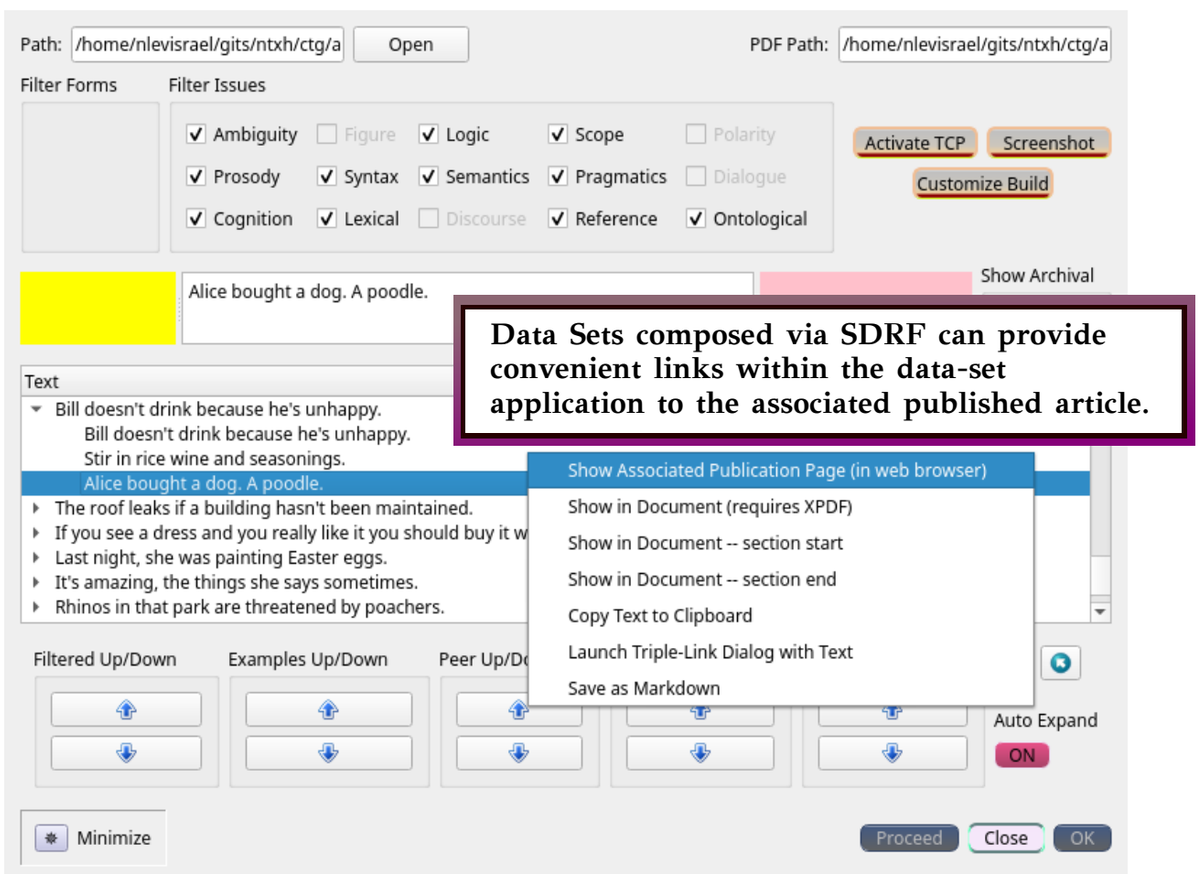

Figure 5: Data Set Linked to Published Article

The data set illustrated Figures 5-7 also demonstrates how our PDF generation tools may be leveraged for data curation. The raw data files and all metadata utilized by the custom data-

set application are extracted from source files composed in a special-purpose PDF-generation language. This source code is marshaled through a multi-faceted data-extraction pipeline which pulls raw data files from document markup (in this case, the raw data encapsulates language samples derived from various linguistic corpora and resources) and also generates machine-readable document encoding, identifying and integrating textual and PDF screen-coordinate locations for sentence and paragraph boundaries and named entities. In short, both raw data and natural-language content is provided through this data set in a structured, machine-readable format that is suitable for multi-publication archiving.

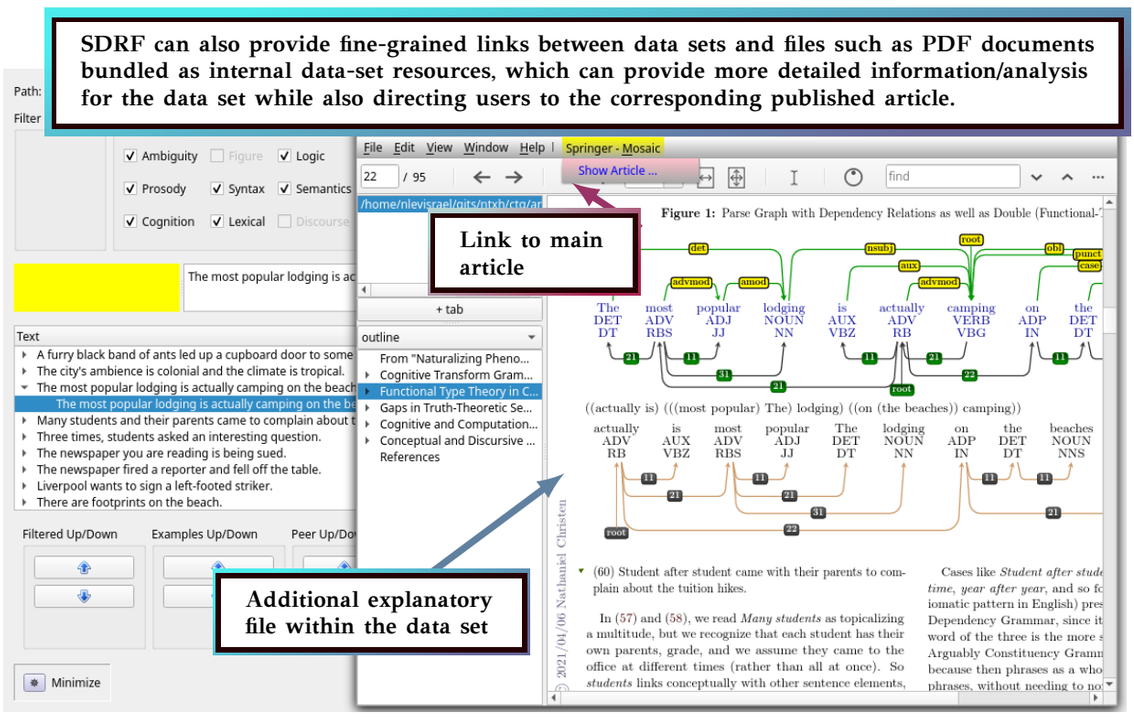

Figure 6: Data Set Linked to PDF File

Collections of publications and data sets curated according to our SDRF protocol can yield archives which augment the value of individual publications and data sets on their own. Such corpora would support searches for keywords, phrases, data types, and code annotations that could simultaneous match both natural-language content (in books and articles) and raw data or data-management code (in data sets). Such corpora could moreover provide a suite of domain-specific data-management and data-visualization software components made available to researchers to assist in their compiling data sets and accompanying code libraries alongside their publications.

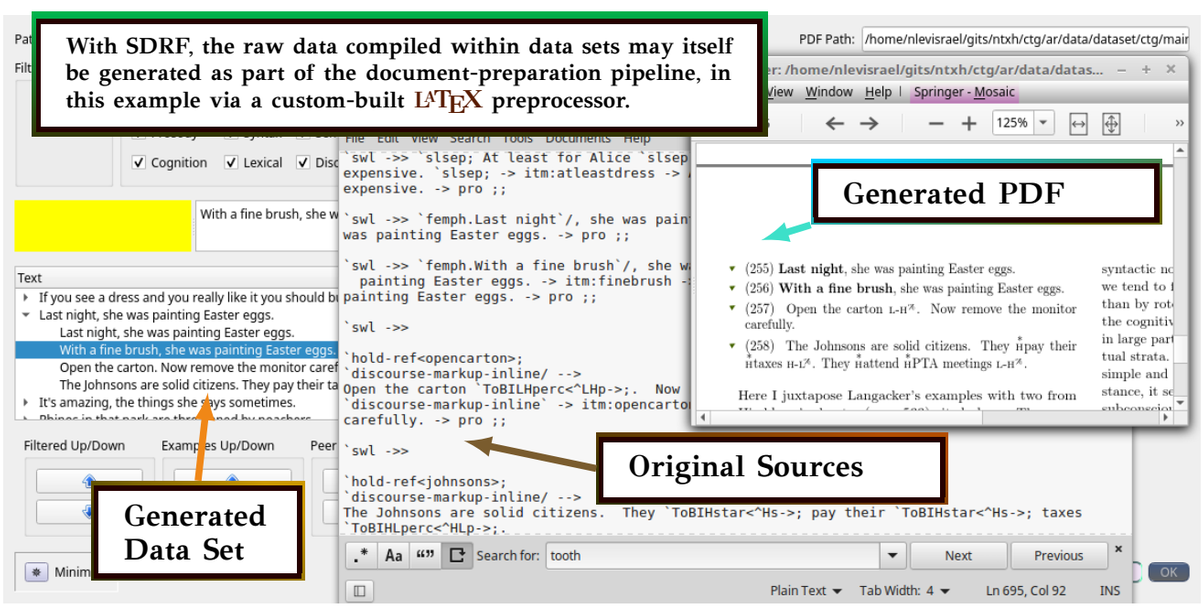

Figure 7: Data Set Extracted from PDF File