



The "Dataset Creator" Database and Document-Generation Framework

Viewing and Creating Datasets and Custom Dataset Applications

Overview

LTS's new Dataset Creator (dsC) framework can be used to (1) generate open-access data sets for publication and (2) curate corpora of academic books/articles. The dsC framework employs innovative text-encoding and hypergraph data-representation protocols that can enhance the query engines and software platforms driving data-hosting and document-sharing repositories. In particular, this technology enables or improves capabilities such as:

- Ensure that data sets are easy to find from publications, and vice-versa.
- Search data sets and publication texts side-by-side.
- Search data sets according to software/computational metadata — e.g., data-type names, data-field names, procedure signatures, serialization formats, visual display formats, and workflow features.
- Publish data sets with customized built-in "dataset applications" for data visualization/analysis, so that researchers are not required to write their own computer code in order to evaluate and reuse the data set (Figures 1-5, below, exemplify problems with uncurated "raw" data sets).
- Provide machine-readable publication texts with precise representations of individual sentences, paragraphs, keywords/phrases, citations/quotations, identifiers for non-textual content (such as figure illustrations and diagrams), and dataset cross-references.
- Integrate machine-readable text with query results, to support search queries such as: "provide the full text for the first sentence in each document containing the search term" or "return the PDF coordinates of the paragraph-start where a figure illustration is referenced."
- Support integrated document and data query APIs via API client libraries and GUI tools.

A screenshot of a digital platform for viewing datasets. On the left, there is a sidebar titled 'Thumbnails' with several small preview images. The main area displays a document titled 'Parkinson's Disease Diagnosis: Effect of Autoencoders to Extract Features from Vocal Characteristics'. Below the title, the author's name 'Ashena Gorgan Mohammadi, Pouy' is listed, followed by the affiliation 'Department of Computer Science, School of Mathem University of Tehran, 14155-6455, Tehran, Iran, Tel: +'. A green button labeled '2' is visible at the bottom of the sidebar. To the right of the main document, there is a large text box with a dark border containing the following text:

This article about Parkinson's Disease is discussed here as a case-study to illustrate problems with (1) data sets that publish raw data without accompanying code and (2) publications linked to multiple data sets in a decentralized manner, forcing researchers to expend time aggregating all data into a central location in order to evaluate or reuse the original authors' research.

Figure 1: Parkinson's Article

Gait in Parkinson's Disease

Jeffrey Hausdorff 

Published: Feb. 25, 2008. Version: 1.0.0

Please include the standard citation for PhysioNet: [\(show more options\)](#)

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* [Online]. 101 (23), pp. e215–e220.

The Parkinson's article references multiple previously-published data sets, including one hosted on PhysioNet, which in turn provides links to raw data files ...



Data Description

Parkinson's disease (PD) is one of the most common movement disorders, affecting approximately 1 million Americans (estimates range between 4 and 6.5 million people worldwide) and about 1% of older adults. In the US alone... 60,000 new cases are

Access

Access Policy:

Anyone can access the files, as

Figure 2: Parkinson's PhysioNet Dataset

Access the files

- [Download the ZIP file \(288.4 MB\)](#)
- Access the files using the Google Cloud Storage Browser [here](#).
- Access the data using the Google Cloud command line tools (p
`gsutil -m -u YOUR_PROJECT_ID cp -r gs://gaitpdb-1.0`
- Download the files using your terminal: [wget -r -N -c -np https://physionet.org/files/gaitpub/1.0.0/](https://physionet.org/files/gaitpub/1.0.0/)

Unfortunately, researchers are not provided with tools to view or analyze this raw data; instead, reusing the original data set requires composing new computer code to convert the raw data into a useful and tractable format.

Folder Navigation: <base>

Name	Size	Modified
GaCo01_01.txt	 1.1 MB	2004-11-14
GaCo02_01.txt	 1.0 MB	2004-11-14
GaCo02_02.txt	 1.0 MB	2004-11-14

Figure 3: PhysioNet Downloads

Solutions to improve academic text-mining and data-sharing are valuable because publishers have invested in document sharing/indexing platforms (such as SpringerNature, Mendeley, CrossRef, ArXiV, JSTOR, or SciVerse) with the recognition that sophisticated data hosting and document curation capabilities raise the profile of their publications: researchers gravitate toward the most full-featured research platforms, and therefore toward the publications hosted on such platforms.

Indeed, data and document repositories which are full-fledged computational ecosystems — and not merely resource collections — are more likely to become scientists' preferred research tools.

However, contemporary publishing platforms have only rudimentary support for standards and initiatives such as Research Objects and **FAIRSHARING** (Findable, Accessible, Interoperable, Reusable) or the guidelines advanced by funding sources such as the Bill and Melinda Gates Foundation and



Raw data files, such as those in this screenshot, cannot be used directly by subsequent researchers; instead, they need to be supplemented with computer code to read the raw data and convert to information that can be visualized and analyzed.

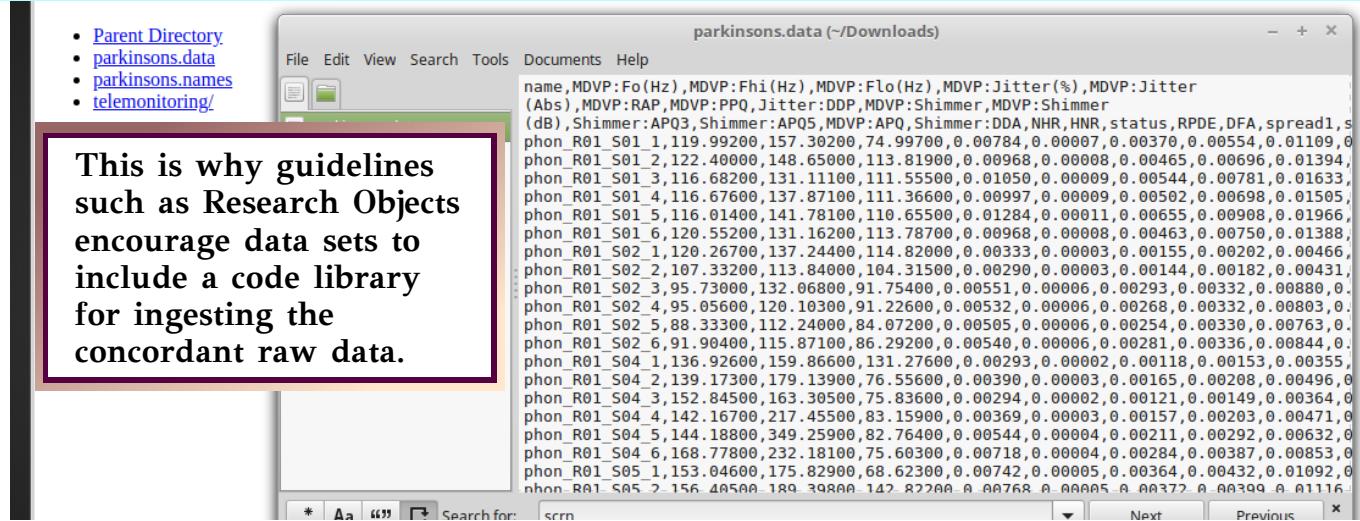


Figure 4: Raw Data is not Useful by Itself

the Chan-Zuckerberg “Invest in Open Infrastructure” program. Tools such as the dsC framework can help publishing platforms realize contemporary standards more effectively, thereby potentially enhancing the prestige, visibility, and popularity of the underlying platform.

The Need for Integrated Data-Hosting and Document Curation

At present, publications (such as academic books or articles) and open-access data sets occupy two largely disconnected ecosystems, where data-hosting services (such as Dryad, Dataverse, OpenScience, or DSpace) are separate from publishers’ portals (where books and articles are listed). As a result of such separation, it is impossible to jointly search publications and data sets using a common query engine or API. Moreover, it can be a tedious process for researchers to actually locate and examine data sets analyzed in scientific literature.

Typical problems arising from data sets being disconnected from publications are demonstrated in Figure 5 (a case-study derived from an article on how background noise affects perceptions of speech clarity — Jan Holub, *et. al.*, “Subjective Speech Quality Measurement Repeatability: Comparison of laboratory test results,” *International Journal of Speech Technology*, 2017); and in Figures 6-9, which show screenshots of an article studying speech, gait, and bioimage markers for Parkinson’s disease (under review for the International Journal of Speech Technology). As this Parkinson’s article considers different biomarker modalities and therefore studies multiple data sets, it furnishes a good illustration of how the lack of integrated data and publication hosting hinders the research process.

Both of these example articles are linked to multiple data sets in a decentralized fashion, with data distributed across numerous platforms (including dropbox, protocols.io, GitHub, Google Cloud,

PLOSOne, Wikimedia, ArXiV, and Figshare) as well as academic and organizational web sites. In order to reproduce the authors' research, or reuse their methods for new investigations, scientists must retrieve data from multiple sources, in many cases studying the original article's supplemental materials to find secondary literature documenting where the relevant data can be found.

In this speech-technology data set, the raw data (audio) files (1) are hosted on a separate platform from metadata files (2). Also, the raw data file set is published just as a zipped folder, with no process to browse the files other than opening them one at a time. In combination, these factors present time-consuming obstacles to data findability, accessibility, interoperability, and reusability, as defined by the FAIRSHARING initiative.

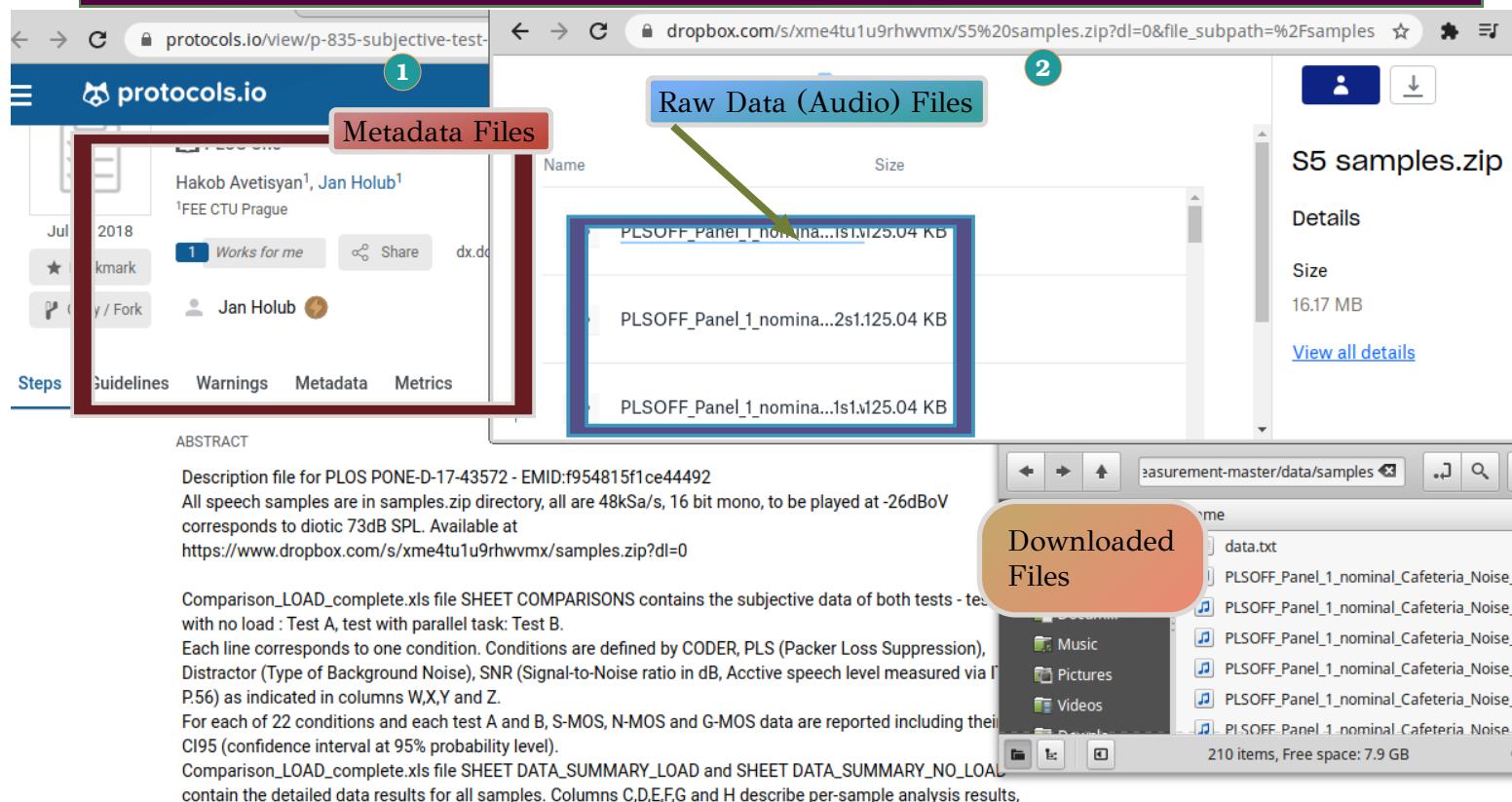


Figure 5: A Case-Study Showing the Limitations of Decentralized Data Hosting

In addition to being a time-consuming bottleneck for research, such ad-hoc data networking makes it difficult for multiple data sets to be merged in a rigorous fashion: this manner of decentralized data-curation leads to data mishmashes with limited scientific integrity, and no reusable digital assets for merging the data into a common format or a "Common Data Model."

Alongside these problem of data decentralization, the majority of published data sets are currently also limited in that they supply only raw data files — without including code to make their raw data accessible to subsequent researchers, either through visual tools or through ingest/analysis algorithms. This forces researchers to write their own code in order to access the data, which can result in unnecessary wasted duplicate efforts, as well as a lack of consistency in how the original data is reused.

Files

Total uncompressed size: 288.4 MB.

Access the files

- Download the ZIP file (288.4 MB)
- Access the files using the Google Cloud Storage Browser [here](#). Login with a Google account is required.
- Access the data using Google Cloud "gsutil":

```
gsutil -m cp -r gs://gaitpdb-1.0.0.physionet.org DESTINATION
```
- Download the files using your terminal:

```
wget -r -N -c -np https://physionet.org/files/gaitpdb/1.0.0/
```

Folder Navigation: <base>			
Name		Size	Modified
GaCo01_01.txt		1.1 MB	2004-11-14
GaCo02_01.txt	←	1.0 MB	2004-11-14
GaCo02_02.txt		1.0 MB	2004-11-14
GaCo03_01.txt		1.1 MB	2004-11-14
GaCo03_02.txt		1.1 MB	2004-11-14

Some information in PhysioNet is encoded in file names, where the initial two letter-pairs and following two number-pairs all provide information about the patient and data source. Unfortunately, encoding data in this manner requires extra computer code when reusing the data base, because the file names need to be analyzed so as to parse the information expressed via the naming conventions.

Figure 6: 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 feature [1] range.

Table 1: A summary of reviewed datasets.

Data type	Description	Study
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 with 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 people 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 Mark Little of the University of Oxford, in collaboration with the National Centre for Voice and 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 measurement.	[15], [17]

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

Figure 7: Table Listing Analyzed Data Sets in the Parkinson's Article



- [8] Sakar, C. Okan, Gorkem Serbes, Aysegul Gunduz, Hunkar C. Tunc, Hatice Nizam, Betul Erdogan 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. Moretti, 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.2013.11.017.
- [10] "Gait in Parkinson's Disease." Gait in Parkinson's Disease V1.0.0. February 25, 2008. <https://physionet.org/content/gaitpdv1.0.0/>
- [11] Rehman, Rana Zia Ur, Silvia Del Di, 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/s41598-019-45267-w.
- [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, 2011.
- [13] UCI Machine Learning Repository: Parkinson's Disease Classification Data Set. <https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification>.
- [14] Sriram, Tarigoppula & Rao, M. & Narayana, G & Vital, T. & Dowdun, M. & Kumar SVGK, (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] Erdemli, Cukur, Ratal et al. "Analyzing the effectiveness of visual features in early telediagnosis of

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

Figure 8: Bibliography (With Data Set Hyperrefs) in the Parkinson's Article

Acknowledgments

We thank Lei Li and Yuncheng Zhou, whose dedicated comments helped us to modify the organization of the manuscript and improve its quality.

Supplementary Material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnins.2019.00874/full#supplementary-material>

... and some data sets analyzed in the Parkinson's article are only identified in "Supplemental Material" sections of its cited papers.

Figure 9: Indirectly Locating Data Sets from Cited Papers

Building Data Sets with ConceptsDB Dataset Creator

In contrast to the paradigm of decentralized curation depicted above, LTS's Dataset Creator can be utilized to build integrated and reusable open-access data sets. With dsC, authors have tools to bundle data with code according to the Research Object protocol and FAIRSHARING guidelines, employing an embedded code base to supply a custom desktop-style dataset application that offers a convenient visual, interactive, and analytic interface to the raw data. The ConceptsDB database engine, which LTS has designed alongside dsC, can optionally be used to update/manage dsC data. ConceptsDB and/or dsC (which may be used together or separately) afford several benefits to publishers, including:

1. Dataset Applications can be configured to link back to the published article, offering an additional route for finding the original article and boosting citations/influence (see Figure 10).
2. Dataset Applications will necessarily be implemented via code libraries encapsulating data access,



libraries which can be leveraged for query engines operating at the data-hosting site, offering more sophisticated and fine-grained query capabilities across multiple data sets. In short, dataset code can be designed to explicate properties of data types, fields, procedures, and file formats, all of which produces queryable metadata that might be leveraged for advanced search features.

3. Documenting dataset applications — via explanations of **GUI** features, compile instructions, test suites, code reviews, and similar semi-formal code representations — can be performed through tools customized for the data-hosting platform, yielding consistent patterns in how data sets are employed, which can benefit human users and also allow some data-acquisition steps to be done automatically.
4. Similarly, dataset applications can expose multiple execution points, thereby supporting workflow description formats. This would allow additional consistency and tooling reuse among multiple data sets sharing a common hosting platform, where workflow representations and implementations could be customized to the platform's architecture and disciplinary foci.
5. Dataset applications can be distributed through platforms such as Qt Creator, Jupyter, SeCo, or ReproZip, yielding publisher-specific tools (e.g., Qt Creator plugins) which can be shared by multiple datasets. This would result in common patterns governing how individual data sets are acquired and accessed, creating a consistency and interoperability among data sets which enhance the value of the overarching hosting platform.

The screenshot shows a user interface for managing dataset issues. At the top, there are two input fields: 'Path: /home/nlevisrael/gits/ntxh/ctg/ar/data/dataset/ctg/samples.nt' and 'PDF Path: /home/nlevisrael/gits/ntxh/ctg/ar/data/dataset/ctg/main.pdf'. Below these are two sections: 'Filter Forms' and 'Filter Issues'.

Filter Forms:

- Text
- Dialog
- Prosodic
- Segment
- Paragraph

Filter Issues:

- Ambiguity
- Roles
- Logic
- Scope
- Types
- Intonation
- Syntax
- Semantics
- Pragmatics
- Dialogue
- Cognition
- Lexical
- Discourse
- Reference
- Ontological
- Idioms
- Figural
- Polarity
- Epistemics
- Context

On the right side, there are several buttons: 'Activate TCP', 'Customize Build', 'More Actions ...', 'Show Archival' (with a dropdown set to 'OFF'), and a large red box containing the text:

Data Sets composed via dsC can provide convenient links within the data-set application to the associated published article.

Below the filter sections, a table displays search results:

	txt	34	sco	22	2
(61)	txt	35	lex	26	2
(62)	txt	36	lex	26	2
	txt	37	ont	26	2
	txt	38	ref	29	2
	dlg	39	ont	29	2
	txt	40	ref	30	2

A context menu is open over the first row of the table, listing options: 'Show Associated Publication Page (in web browser)', 'Show in Document (requires XPDF)', 'Show in Document -- section start', 'Show in Document -- section end', 'Copy Text to Clipboard', 'Launch Triple-Link Dialog with Text', and 'Highlight (scroll from here)'.

At the bottom of the interface are several buttons: 'Minimize', 'Proceed', 'Close', and 'OK'.

Figure 10: Data Set Linked to Published Article

In sum, data-hosting platforms which provide data sets curated in accordance with FAIRsharing and Research Object guidelines are more likely to become go-to resources for researchers and educators, because data sets obtained from such platforms are more likely to be interoperable and reusable.

As shown in Figures 10-14, dataset applications also enhance the value of publications, by creating interactive software modules where ideas explored in publication texts are examined or demonstrated using digital assets. Over time, such dataset applications could potentially be reused for different research projects and evolve into general-purpose software tools for the relevant scientific field.

Custom Dataset Applications can provide interactive features specific to the kind of data being shared. This example shows a data set comprised of 553 language samples analyzed from the perspective of Cognitive Grammar. The samples are organized into groups discussed together for purposes of exposition, and also indexed and filtered in terms of thematic issues and discursive format, and users can examine and navigate through the sample collection via groups, issues, formats, or by reading accompanying text discussing the samples in depth.

Format Filters

Filter Forms: Text, Dialog, Prosodic, Segment, Paragraph

Filter Issues: Ambiguity, Roles, Logic, Scope, Types, Intonation, Syntax, Semantics, Pragmatics, Dialogue, Cognition, Lexical, Discourse, Reference, Ontological, Idioms, Figural, Polarity, Epistemics, Context

Text area: Time after time, tourists walk by this building with no idea of its history.

Highlighted Sample: Time after time, tourists walk by this building with no idea of its history.

Operating System Profile: Linux (Generic), 32 Bit, 64 Bit

Compile Options: Use UDPipe, Build R/Z (scripting), Gen Test, Use XPDF, Build External XPDF Application, Build PDF Scraper, Build DGDB Components, Build Research Object Information Console.

Select User Role: User, Reader, Researcher (Default), Author, Editor, Tester, Administrator.

Configuration Console for Application Install: Proceed, Cancel, OK

Show Archival: ON

Options to Control Details about How Samples are Displayed: ON

Group and Sample IDs:

Form	#	Issue	Page	Section
txt	13	log	13	1
txt	14	sco	13	1
	(39)			
	(40)			
	(41)			
	(42)			
txt	15	log	16	1
txt	24	sem	17	1
txt	29	sem	19	2
txt	31	sem	20	2

Page and Section Refs: First, Auto Expand ON, Proceed, Close, OK

Figure 11: Features of a Custom Dataset Application



Additional features of this dataset application are **GUI** tools to help construct parse graphs and structural representations of sentence structure, which can then be displayed as text or figures within accompanying publications. Although the data structures and **GUI** classes shown here are specific to linguistics, dataset applications in other disciplines may similarly allow figures or diagrams defined in publications to be parsed and included in an accompanying data set.

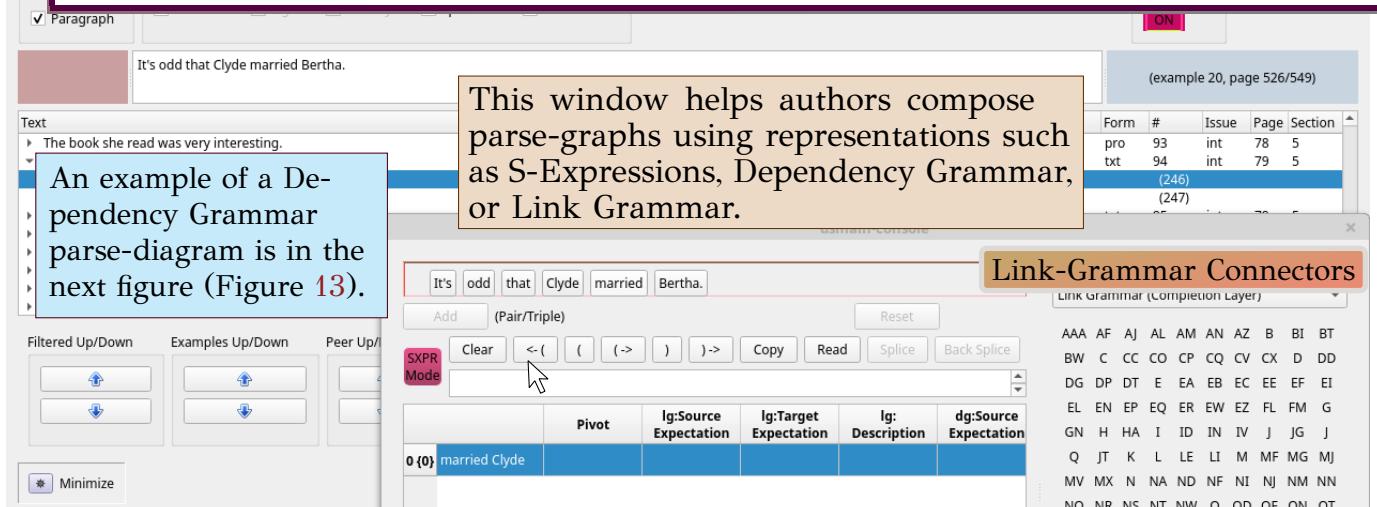


Figure 12: Additional Interactive Features of a Custom Dataset Application

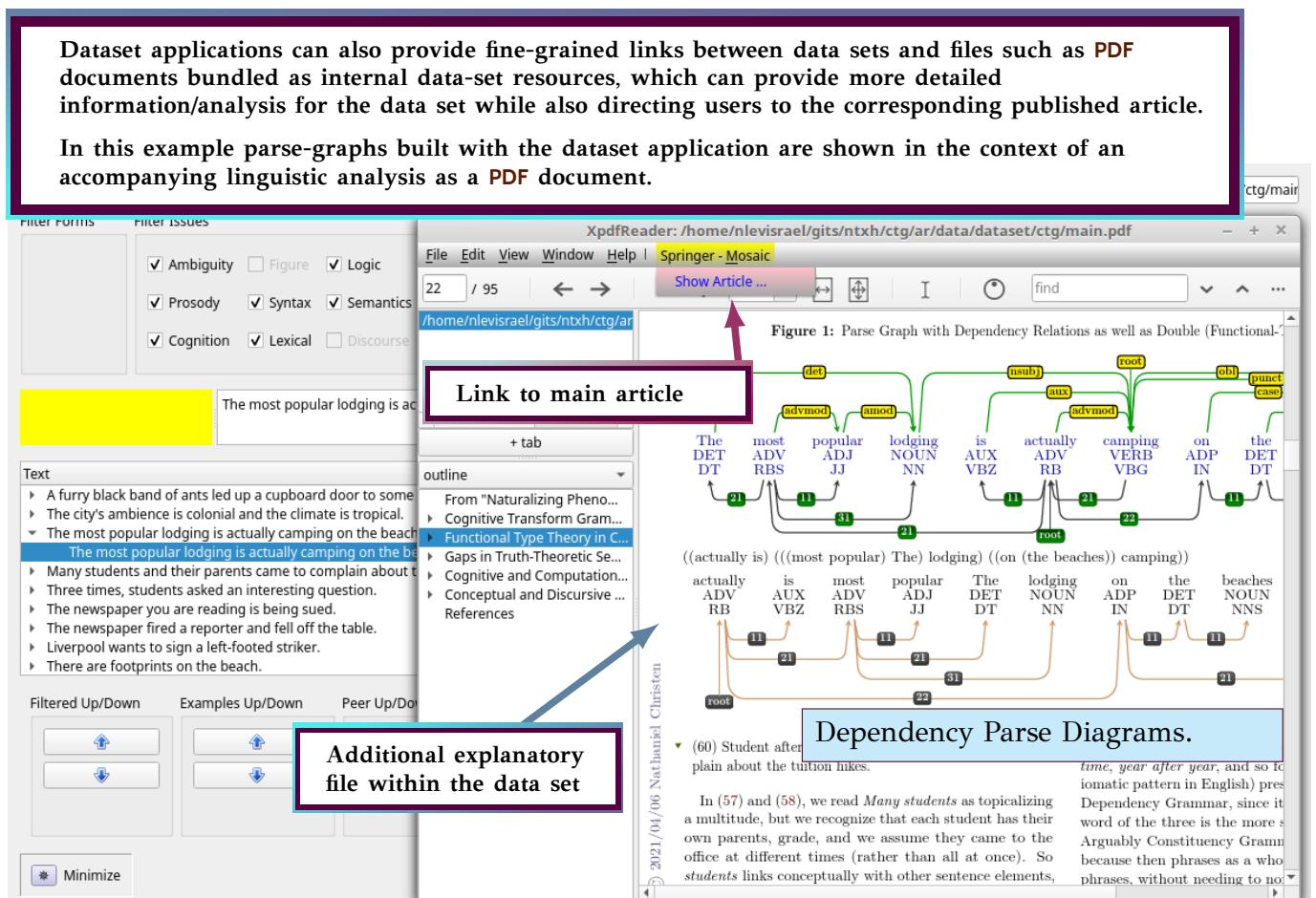


Figure 13: Data Set Linked to PDF File

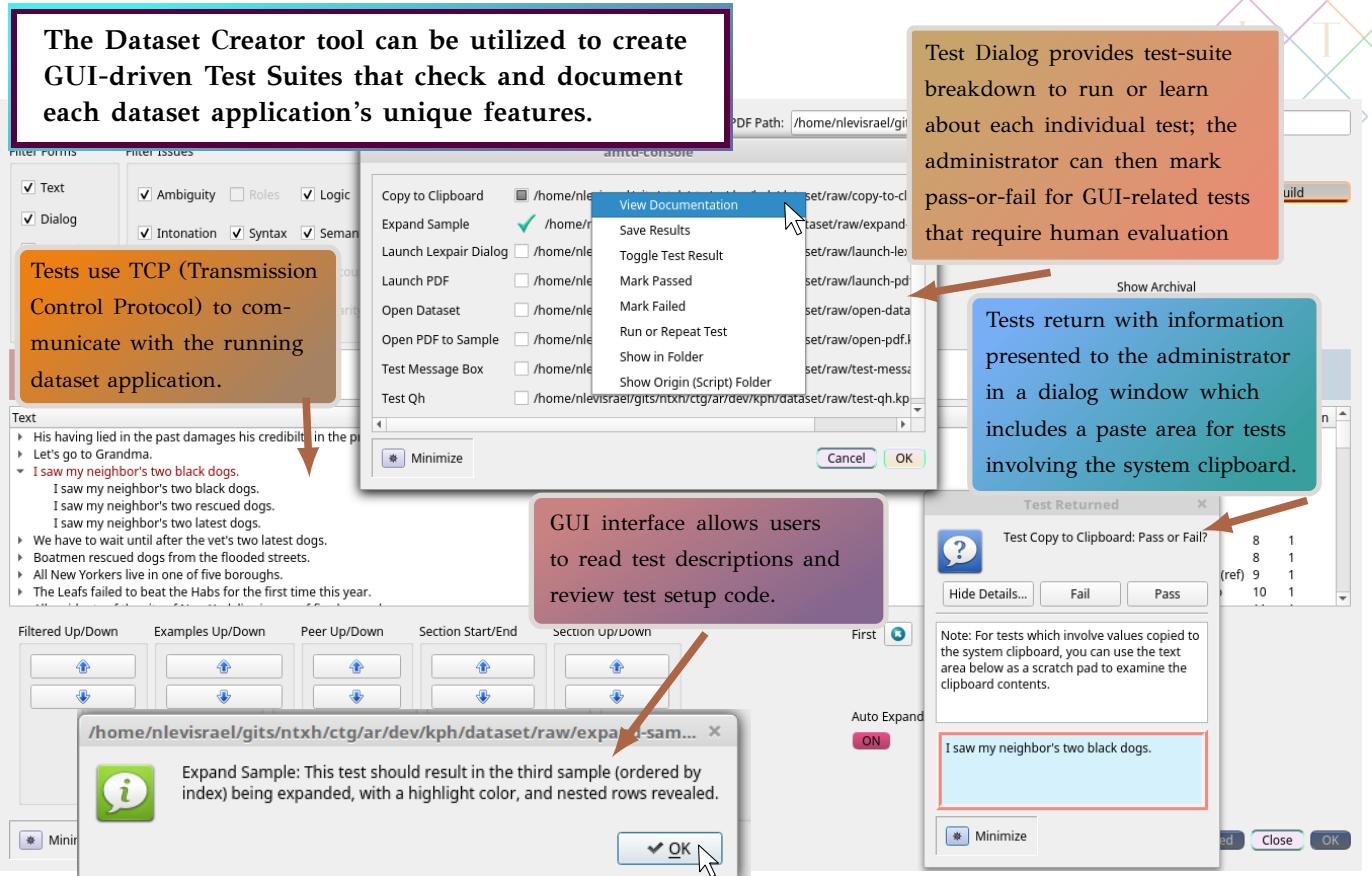


Figure 14: Dataset Application GUI Test Suite

With custom dataset applications, the raw data compiled within data sets may itself be generated as part of the document-preparation pipeline, in this example via a custom-built **LATEX** preprocessor.

Generated Data Set

Original Sources

Generated PDF

PDF Path: /home/nlevisrael/gits/ntxh/ctg/ar/data/dataset/ctg/main.pdf

XpdfReader: /home/nlevisrael/gits/ntxh/ctg/ar/data/dataset/ctg/main.pdf

File Edit View Tools Documents Help

81 / 95

(255) Last night, she was painting Easter eggs.
(256) With a fine brush, she was painting Easter eggs.
(257) Open the carton L-H%. Now remove the monitor carefully.
(258) The Johnsons are solid citizens. They *ipay their *taxes H-L%. They hattend *PTA meetings L-H%.

Here I juxtapose Langacker's examples with two from

Figure 15: Data Set Extracted from PDF-Generator File

Generating Machine-Readable Publication Text with HTXN



ConceptsDB and dsC employ a novel Hypergraph Text Encoding Protocol (HTXN) for text storage/representation. The HTXN protocol can also be used independently of the dsC framework for general-purpose document generation. HTXN can output **HTML**, **LATEX**, and **PDF** files while also producing machine-readable text representations and, concurrently, databases suitable for text mining.

The HTXN document-generation pipeline covers multiple stages which aggregate document text and metadata into a queryable resource that recognizes both pure-text content (documents considered as word/sentence/paragraph sequences) and as readable **PDF** files (where sentence and paragraph boundaries are mapped to **PDF** page and screen coordinates). Text and **PDF**-coordinate representations are generated in stages both before and during **PDF** generation, yielding **LATEX** auxiliary files which are merged by post-processing algorithms to create an integrated, queryable document model, called a "Semantic Document InfoSet" (**SDI**) (see Figure 4). This **SDI** is then grouped into pages and may be loaded by a special **PDF** viewer as each page is shown in the **PDF** window. The **SDI** can then be further annotated to create microcitation links between publications and data sets.

The **SDI** information from many documents could also be deposited in a common database to create an advanced full-text query feature for publication corpora. A detailed overview of HTXN and the **SDI** generation algorithms and database persistence options will be provided by LTS on request.

Documents generated from HTXN files go through several stages of **LATEX** and **PDF** processing (1) and (2) to identify sentence/paragraph boundaries and other important document landmarks, such as citations and proper names (3), as well as annotations referring to an accompanying data set if present (4).

After Compilation the Full InfoSet can be Subdivided by Publication Page

introduction.gtex (~/gits/ntxh/ctg/dev/documents/ctg/intro.gtex)

```
&type DGH_SDI_Paragraph {1}
:i:1 :j:2 :p:3 :s:4 :e:5 :x:6 :y:7
:ex:8 :ey:9 :o:10 :f:11 :d:12 ;

&type DGH_SDI_Sentence {10}
:i:1 :p:2 :s:3 :e:4 :x:5 :y:6
:ex:7 :ey:8 :o:9 :d:10 ;
```

1 Sentence/Paragraph Info (pre- and post- pdfLatex cycle).

intro.gtex (~/gits/ntxh/ctg/dev/documents/ctg/intro.gtex)

```
&type GH_SDI_Paragraph {4}
:i:1 :j:2 :s:3 :e:4 ;
&/
!/_ DGH_SDI_Paragraph
:$: 1
$j: intro.gt
$p: 1
$#: 2
$e: 1307
$x: 32
$y: 211
$ex: 205
```

2

introduction.gtex (~/gits/ntxh/ctg/dev/documents/ctg/intro.gtex)

```
File Edit View Search Tools Documents Help
read
immediately =\_em
= midsep
===
== ref:harrycause
This
time
Harry =\_em\_.
didn't
cause
our
defeat
===
:opercarton Open the carton^LHp-. Now remove the monitor carefully.
===
:johnsons The Johnsons are solid citizens. | They^Hs-pay their^Hs-taxes^Hlp-.
They^Hs-attend^Hs-PTA meetings^LHp-.
```

3 Citations and Named Entities

intro.gtex (~/gits/ntxh/ctg/dev/documents/ctg/intro.gtex)

```
$& BittnerSmithDonnelly %% (1:main) 2823 3113
$& BlackwellPragmatics %% (1:main) 13680 13701
$& ErwanMoreau %% (1:main) 12193 12194
$& GoertzelPLN %% (1:main) 12191 12192
$& HolmqvistDiss %% (1:main) 2403 2404
$& InteractingConceptualSpaces %% (1:main) 2714 2822
$& JeanPetitot %% (1:main) 3114 3137
$& JordanZlatev %% (1:main) 3318 3338
$& KennethHolmqvist %% (1:main) 2383 2402
$& MattSelway %% (1:main) 2298 2382
$& OlavWiegand %% (1:main) 3138 3315
$& RaubalAdams %% (1:main) 2546 2709
$& RaubalAdamsCSML %% (1:main) 2710 2711
$& Schneider %% (1:main) 12046 12188
$& SleatorTamerley %% (1:main) 12189 12190
$& TerryRegier %% (1:main) 2405 2482
$& WiegandGestals %% (1:main) 3316 3317
$& Zenker %% (1:main) 2712 2713
```

= 2371:
= 1
Matt Selway

4 Dataset Annotations Extracted from the Associated Publication

Figure 16: Document-Generation and Machine-Readable Fulltext Derived from HTXN