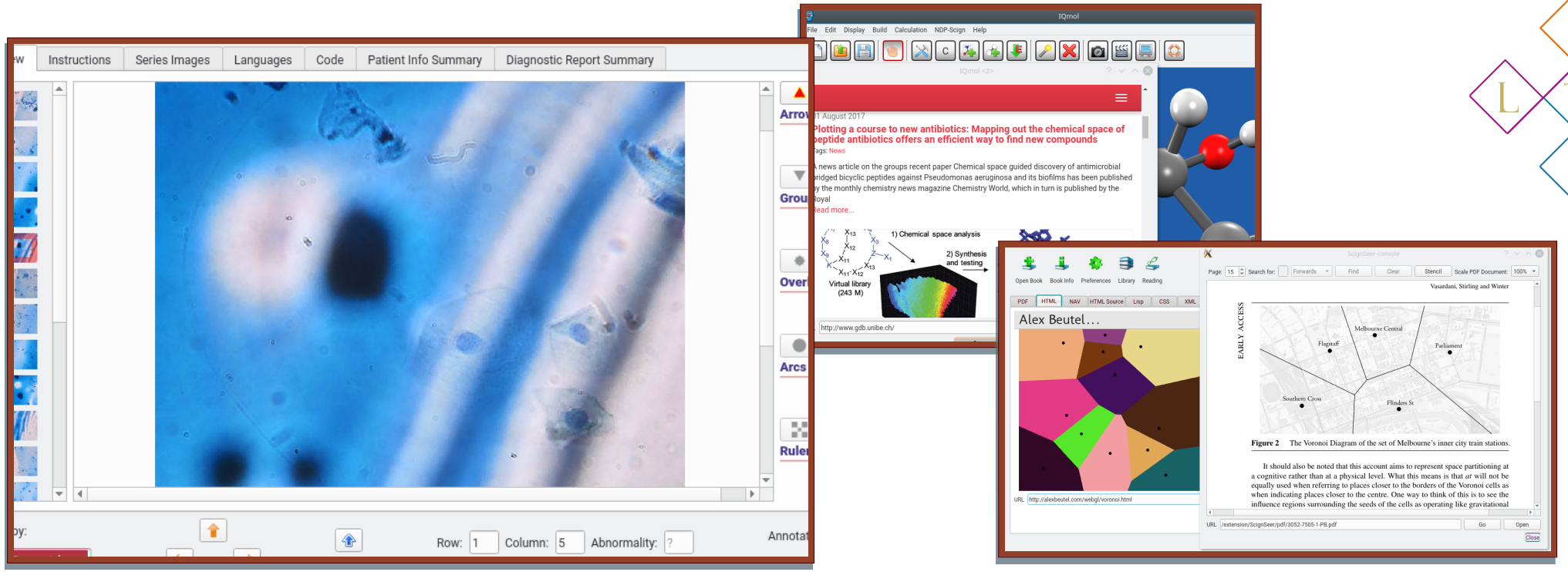
Dataset Creator

("dsC")



Linguistic Technology Systems (LTS)
Amy Neustein, Ph.D., Founder and CEO
amy.neustein@verizon.net
(917) 817-2184

Linguistic Technology Systems

Dataset Creator and Dataset Applications

Dataset Creator Defined

- Dataset Creator (dsC) is a collection of code libraries and project templates which allow authors and publishers to implement interactive software customized for individual research data sets.
- With Dataset Creator, code and data can be unified into a Research Object bundle providing a self-contained "Dataset Application" for responsive, desktop-style exploration of each published data set (see slides 4-21 for Dataset Application case-studies and screenshots).

Features of Dataset Applications

Native applications offer superior User Experience, leveraging distinct interactive features of desktop GUIs: context menus, dialog boxes, tool tips, Multiple Window Display, dock windows, etc:

- Readers of scientific literature can navigate through data sets with GUI controls optimized for the structure and parameters of each data set, synced with research publications (books, articles, and machine-readable manuscripts). This helps researchers to understand the derivation of research findings (particularly when trying to replicate these findings).
- With dsC, Scientists can also build innovative data-collection instruments alongside interactive Research Object applications, providing useful software at each stage of the research process.

Group 1: Features of Dataset Applications

User Interface Features Typical of Dataset Applications

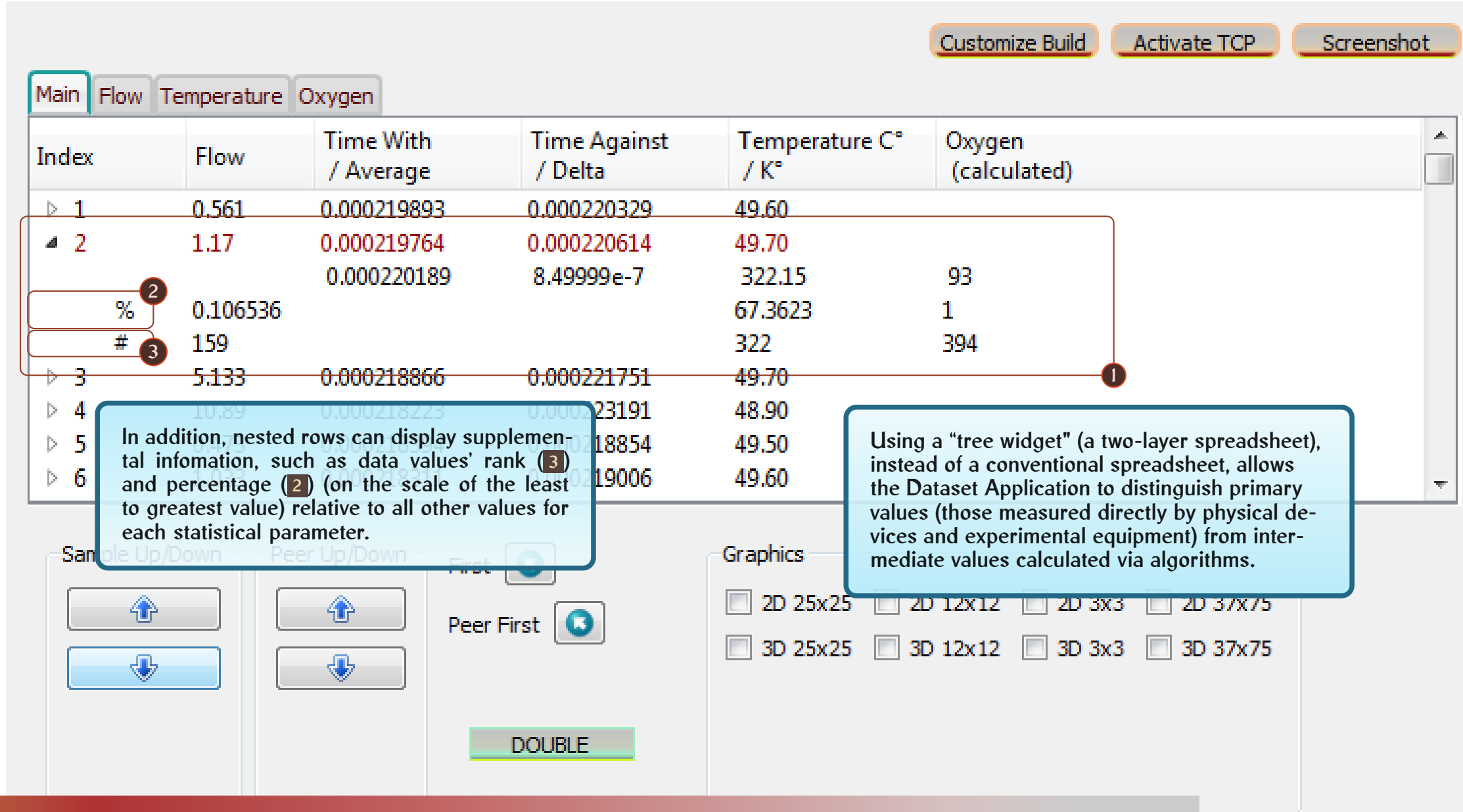
The code for each dsC data set includes a customized "Dataset Application" which displays individual samples and groups of samples via 2D, 3D, and native-compiled GUI controls. Each Dataset Application can thereby make use of advanced visual and interactive features that are uniquely possible when using customized, native-compiled GUI classes. The following screenshots will show several examples of these features, including:

Specialized Top-Level Controls Tree Widgets, Stacked Widgets, and Graphics Scenes.

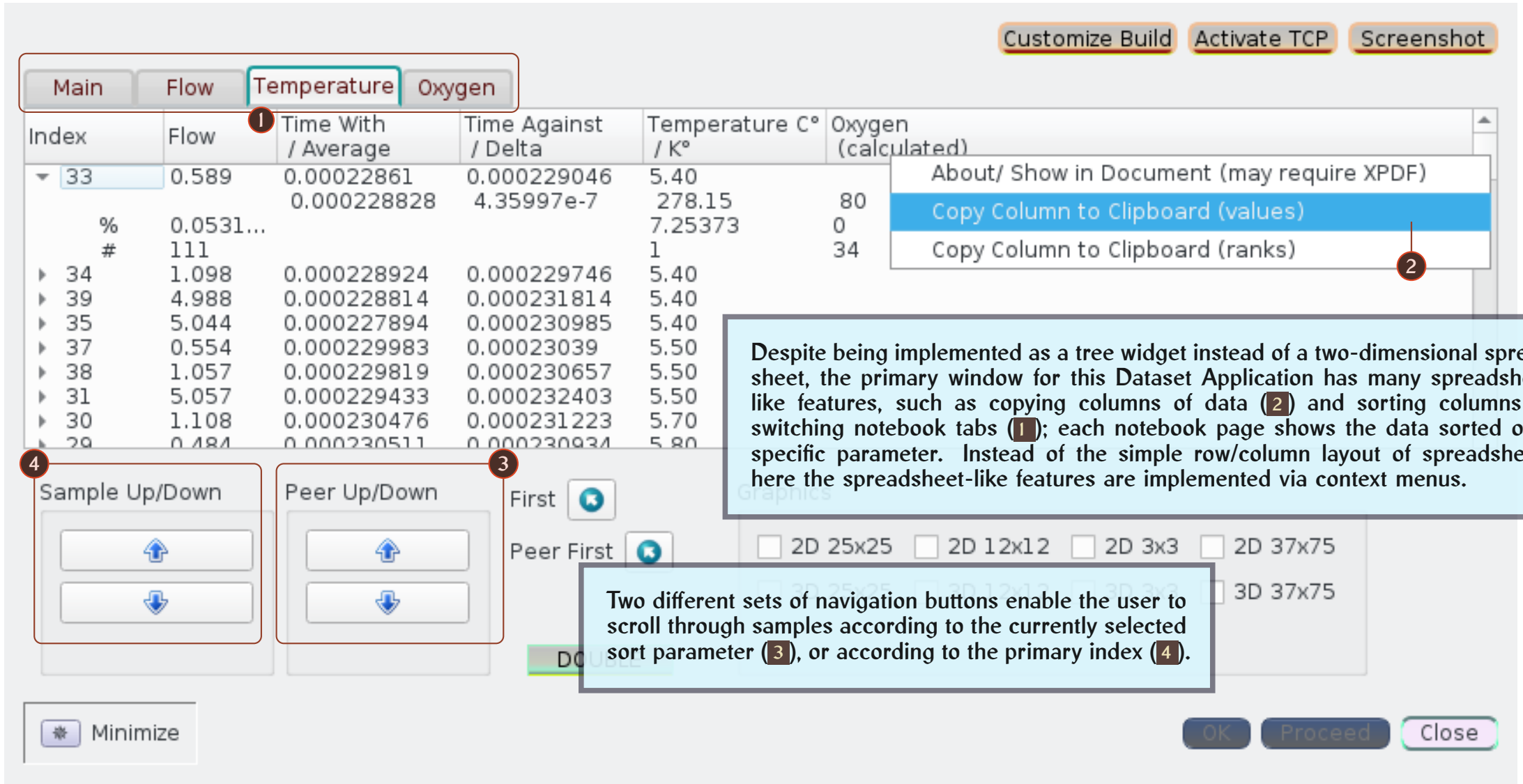
Context Menus Systematically organize functionality around UI layouts.

Multi-Window Displays Divide application functionality in multiple specialized top-level windows and/or dialog boxes.

Initial Application Window



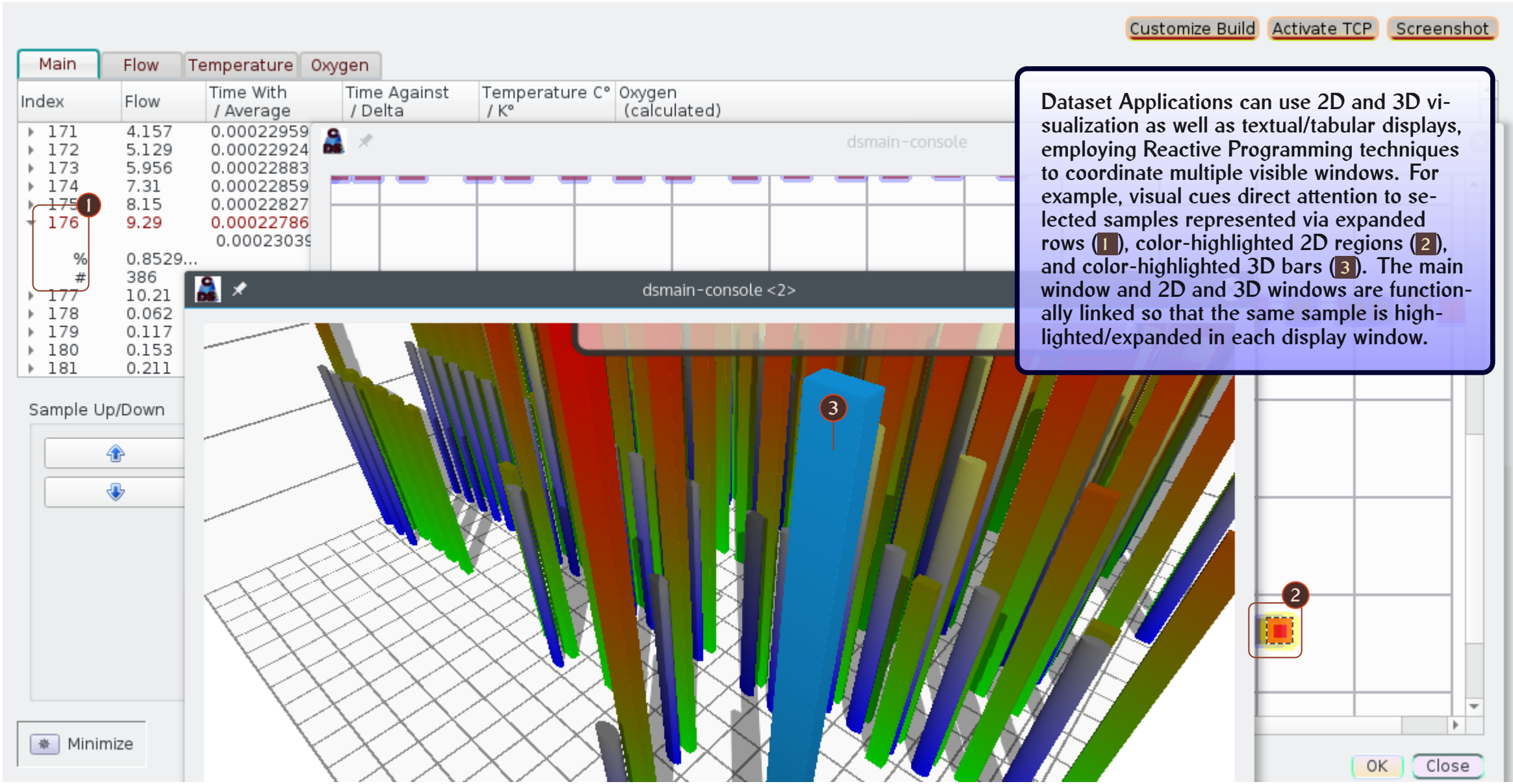
Interacting with the Main Window



Despite being implemented as a tree widget instead of a two-dimensional spreadsheet, the primary window for this Dataset Application has many spreadsheet-like features, such as copying columns of data (2) and sorting columns by switching notebook tabs (1); each notebook page shows the data sorted on a specific parameter. Instead of the simple row/column layout of spreadsheets, here the spreadsheet-like features are implemented via context menus.

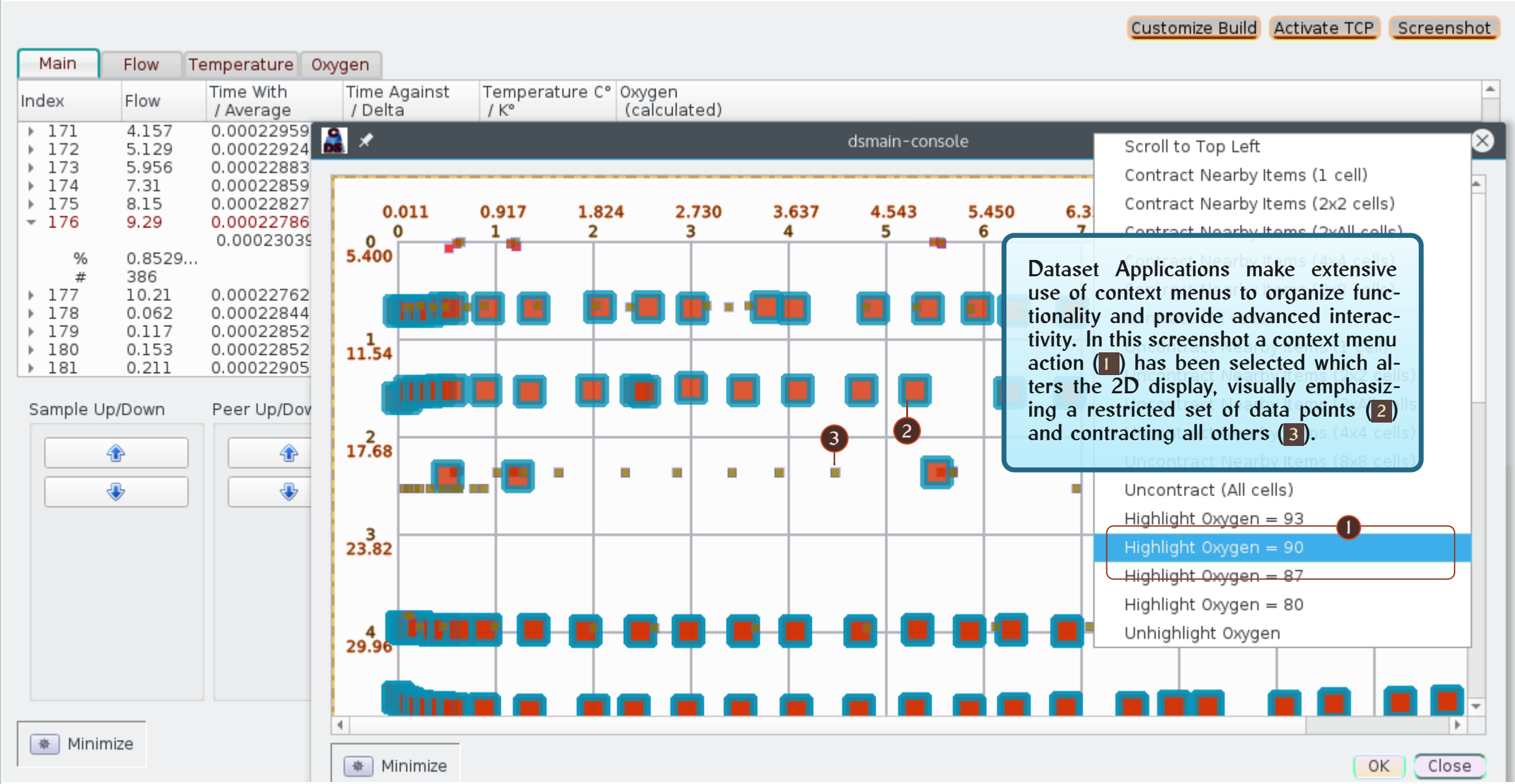
Two different sets of navigation buttons enable the user to scroll through samples according to the currently selected sort parameter (3), or according to the primary index (4).

Coordinated Data Visualization



Dataset Applications can use 2D and 3D visualization as well as textual/tabular displays, employing Reactive Programming techniques to coordinate multiple visible windows. For example, visual cues direct attention to selected samples represented via expanded rows (1), color-highlighted 2D regions (2), and color-highlighted 3D bars (3). The main window and 2D and 3D windows are functionally linked so that the same sample is highlighted/expanded in each display window.

Interacting with the Visuals



Dataset Applications make extensive use of context menus to organize functionality and provide advanced interactivity. In this screenshot a context menu action (1) has been selected which alters the 2D display, visually emphasizing a restricted set of data points (2) and contracting all others (3).

Using Dataset Applications to Understand Experimental Protocols and Research Methods

Obtaining Information About Modeling Parameters

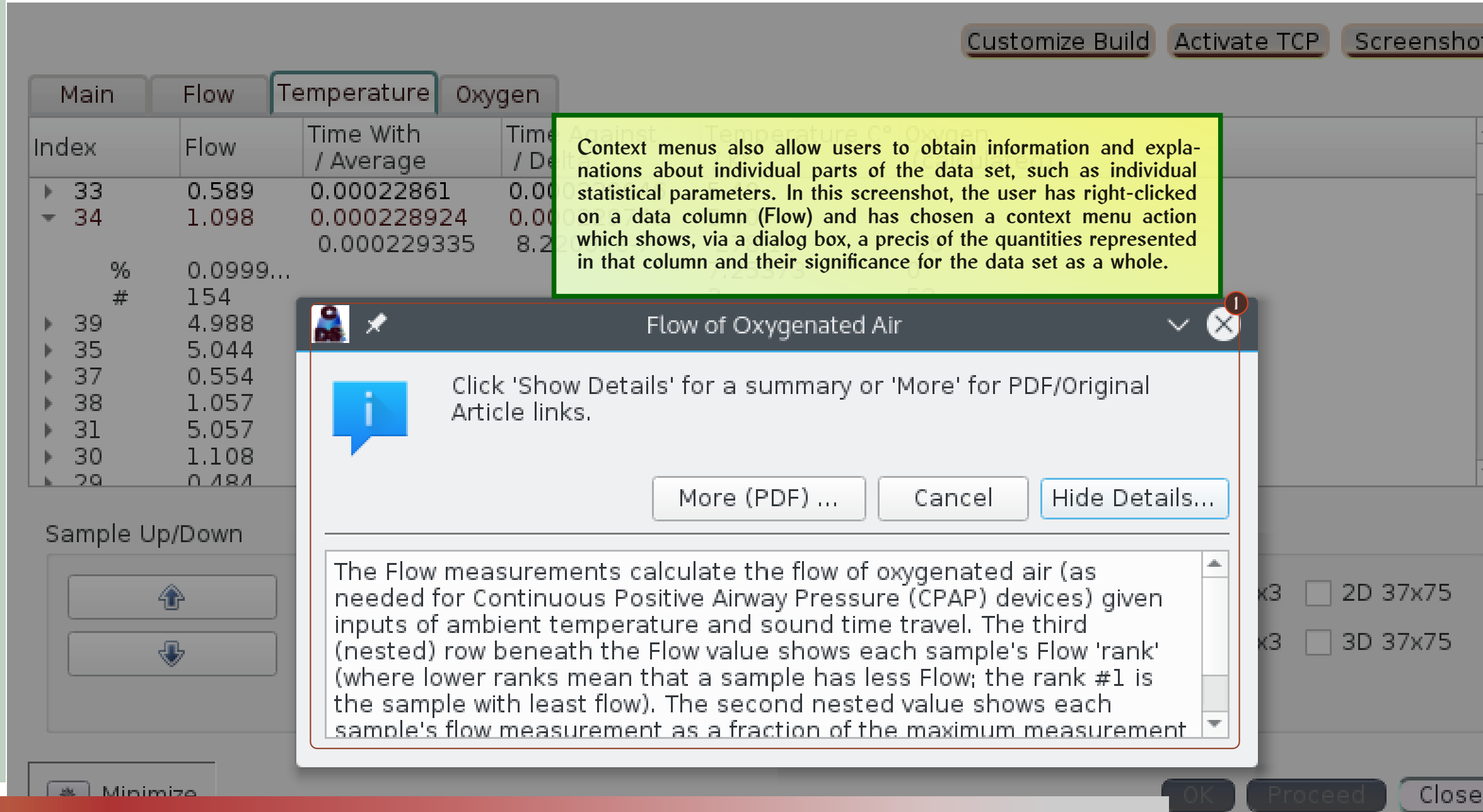
In addition to interactive visualization, Dataset Applications are useful pedagogic tools. Within Dataset Applications, modeling units such as statistical parameters and record fields are visible in situ within a GUI — identified by labels, buttons, and other interactive micro-controls. As a result, users encounter modeling elements in a structured visual-interactive context. To help the reader learn more about modeling elements, Dataset Applications are equipped with several pedagogic features (which are shown on the following slides):

"About" Dialogs Brief summaries of research terms and parameters.

XPDF Links Links back to research articles read in an embedded PDF viewer.

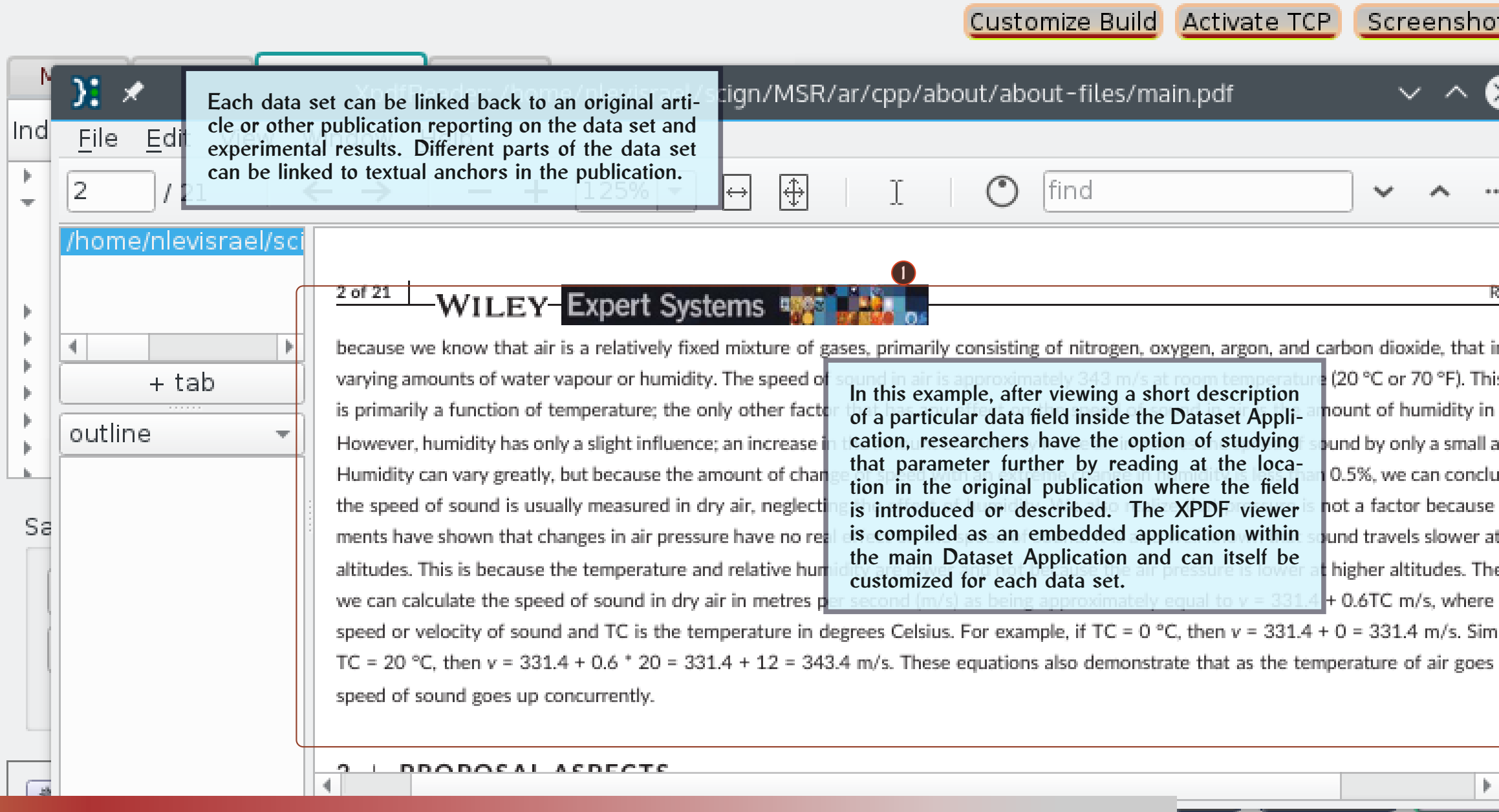
XPDF Enhancements The XPDF viewer can be customized for each data set and included with dataset code, with extra features to integrate article or book texts with Dataset Applications.

Obtaining Information About Parameters



Context menus also allow users to obtain information and explanations about individual parts of the data set, such as individual statistical parameters. In this screenshot, the user has right-clicked on a data column (Flow) and has chosen a context menu action which shows, via a dialog box, a precis of the quantities represented in that column and their significance for the data set as a whole.

Embedding XPDF



Testing and Fine-Tuning Dataset Applicationss

Tools for Editors and Developers

Although ordinary users can explore and visualize dsC data sets "Out of the Box", advanced users have many options for customizing their build of the application in terms of their specific roles and available 3rd-party libraries. These fine-tuning possibilities include:

Test Suites Tools for creating and/or running test suites to ensure that the Dataset Application works across platforms.

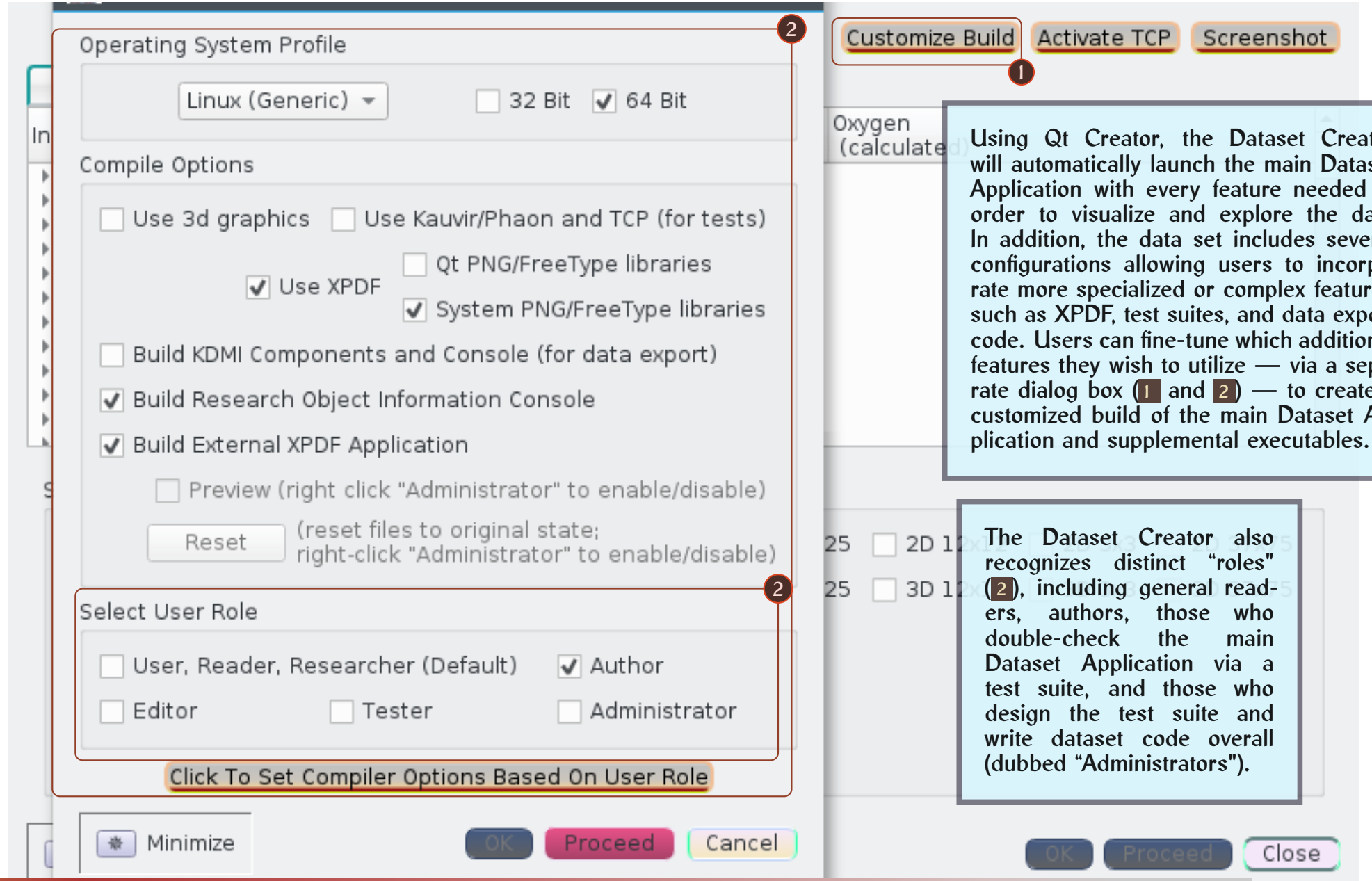
Data Export Tools for reusing data in other projects.

External Libraries Some features like XPDF and 3D graphics require libraries that cannot be published with the data set in source code form. Advanced users can slect which of these libraries to incoporate into thir version of the Dataset Application.

Scripting Data sets can comoile their own scripting environment to automate testing and manipulation of research data.

Networking Dataset Applications can use an embedded TCP server to communicate with other applications, enabling multi-application workflows (this is also how testing is implemented).

Configuring the Data Set Application

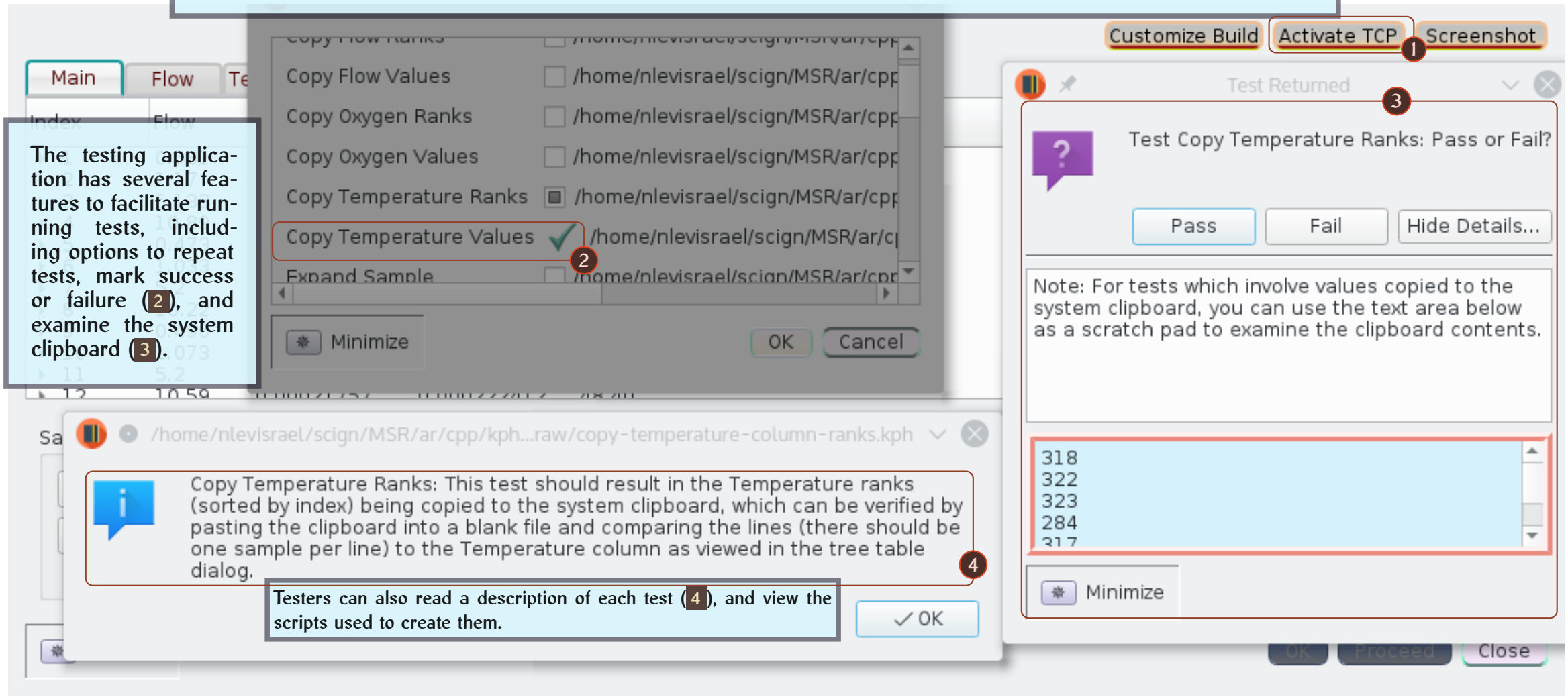


Using Qt Creator, the Dataset Creator will automatically launch the main Dataset Application with every feature needed in order to visualize and explore the data. In addition, the data set includes several configurations allowing users to incorporate more specialized or complex features, such as XPDF, test suites, and data export code. Users can fine-tune which additional features they wish to utilize — via a separate dialog box (1 and 2) — to create a customized build of the main Dataset Application and supplemental executables.

The Dataset Creator also recognizes distinct "roles" (2), including general readers, authors, those who double-check the main Dataset Application via a test suite, and those who design the test suite and write dataset code overall (dubbed "Administrators").

Testing the Data Set Application

Dataset Creator includes a sophisticated framework for building and running test suites to ensure that raw data is processed correctly and that User Interface components work properly on different Operating System platforms. This includes a separate testing application that sends instructions to the main Dataset Application via TCP (1).



The testing application has several features to facilitate running tests, including options to repeat tests, mark success or failure (2), and examine the system clipboard (3).

Testers can also read a description of each test (4), and view the scripts used to create them.

Features of Dataset Applications for Books

Datasets Ccompiled From Book Examples

The remaining screenshots demonstrate how data sets can be used even outside of a lab contxt generating experiment data. The pictured data set represents a corpus of linguistic examples mined from Wiley's *Blackwell Handbook of Pragmatics*. Creating data sets from book-length publications can encompass several steps:

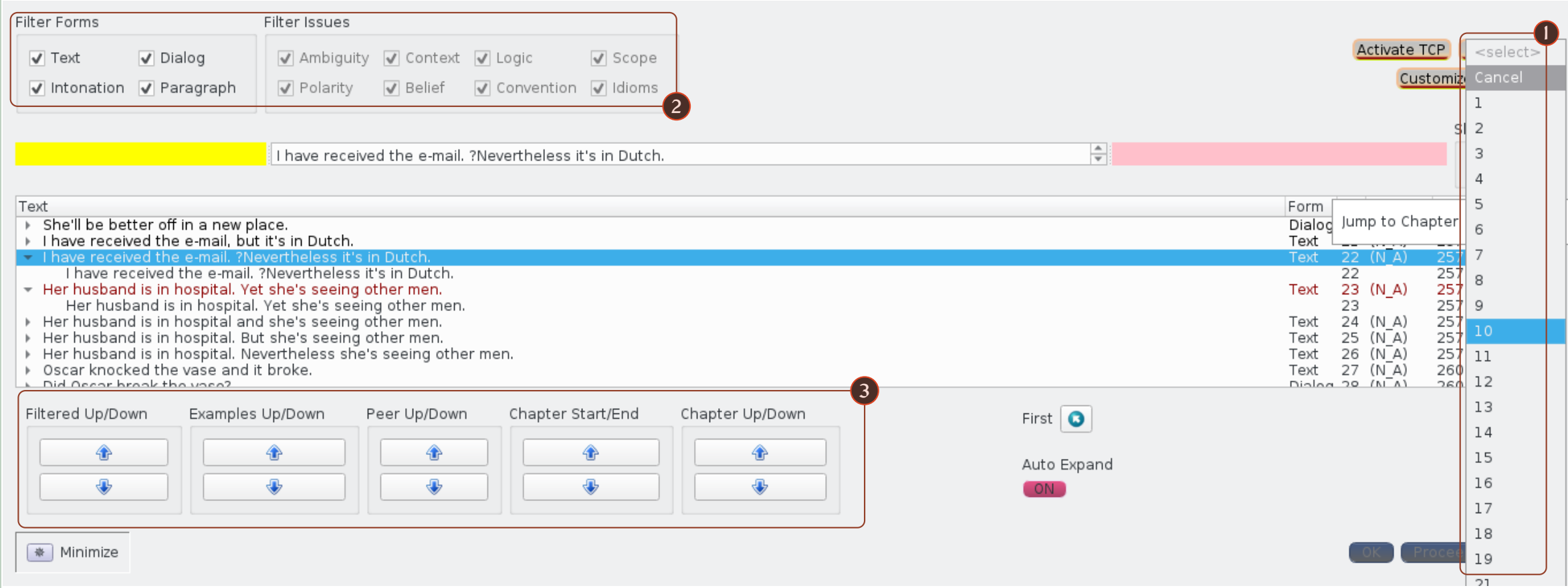
Text Mining In the case of linguistics, this involves locating example sentences within linguistics texts and storing them as an independent corpus.

Canonical Formatting If possible, linguistics texts should be annotated so that extracting exmples can be automated.

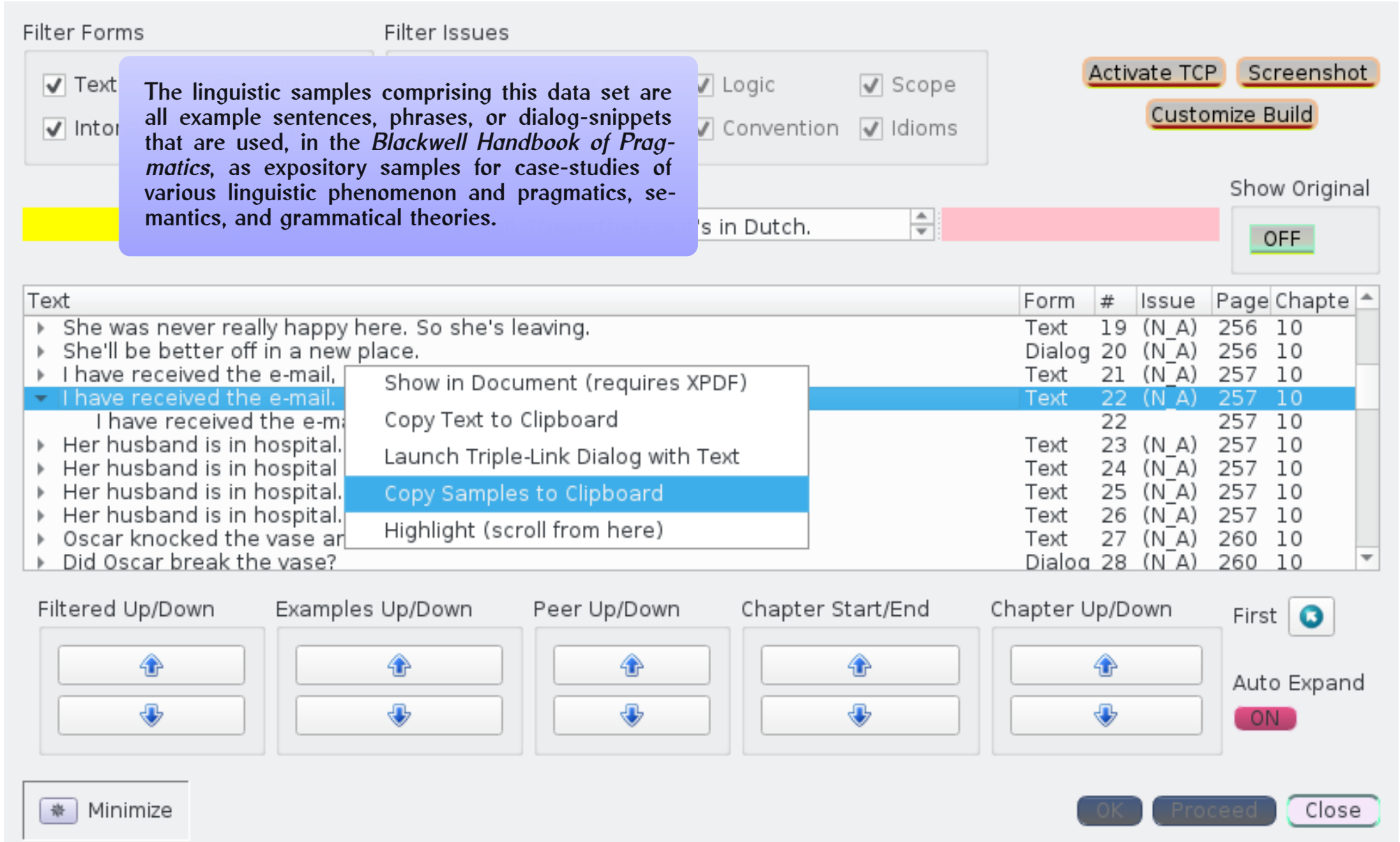
Annotation Linguistic corpuses are often annotated to identify structural details, beyond raw text, in each sample.

Creating a Data Set from a Book

This screenshot shows a linguistics dataset that illustrates several advanced interactive features made possible by the Dataset Creator's Qt-based front-end technology. Useful features include context menus embedded with drop-down selections (1) and button/checkbox groups for filtered scrolling through a list of samples (2 and 3).



Interacting with Data Samples



Linking Back to the Book

Filter Forms

Text Dialog

Intonation Paragraph

Filter Issues

Ambiguity Context

Polarity Belief

In France, Watergate wouldn't

Text

On the table.
Every bottle is empty.
She seized the knife and stabbed her husband.
The Boston Marathon will take place next week. Max thoug
My friends were under the impression that I was running a
Sue believes Luke has a child and that Luke's child will visit
In France, Watergate wouldn't have done Nixon any harm.
In France, Watergate wouldn't have done Nixon any har
In France, Watergate wouldn't have done Nixon any harm b
The crook paid them with fake money.
The crook thought he was paying them with fake money, b
We do not know much about this part of the brain, which p

Filtered Up/Down Examples Up/Down Peer Up/Down

Minimize

XpdfReader: /home/nlevisrael/scign/HP/ar/cpp/about/about-files/main.pdf

File Edit View Window Help

690 / 867 113% find

After browsing through the data set, users can link back to the original text to see the current author's discussion of particular examples.

15. The Pragmatics o...
16. Pragmatics of La...
17. Constraints on Ell...
III Pragmatics and its Int...
18. Some Interaction...
19. Pragmatics and A...
20. Pragmatics and S...
21. Pragmatics and t...
22. Pragmatics and t...
23. Pragmatics and I...
24. Historical Pragma...
25. Pragmatics and L...
26. Pragmatics and C...
IV Pragmatics and Cogni...
27. Relevance Theory
28. Relevance Theory...
29. Pragmatics and C...
30. Pragmatic Aspect...
31. The Pragmatics o...
32. Abduction in Nat...
Bibliography
Index

from the matched spaces to create a **blended mental space** with emergent structure. This creates a conceptual integration network of the form shown in figure 29.4. The generic space represents the structure shared by the inputs. The square in the blended space stands here for the emergent structure which arises in the blending.

So, for example, one way to understand the counterfactual in (6):

(6) In France, Watergate wouldn't have done Nixon any harm.

is to build a conceptual integration network that partially matches two input spaces with prominent aspects of the American political system and the French political system, respectively, and develops an emergent blended space

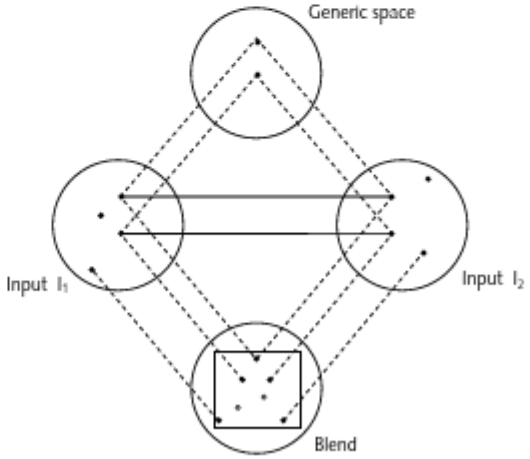


Figure 29.4 Diagram showing conceptual blending

A Linguistics Annotation System

Tools to Facilitate Annotating Linguistic Corpora

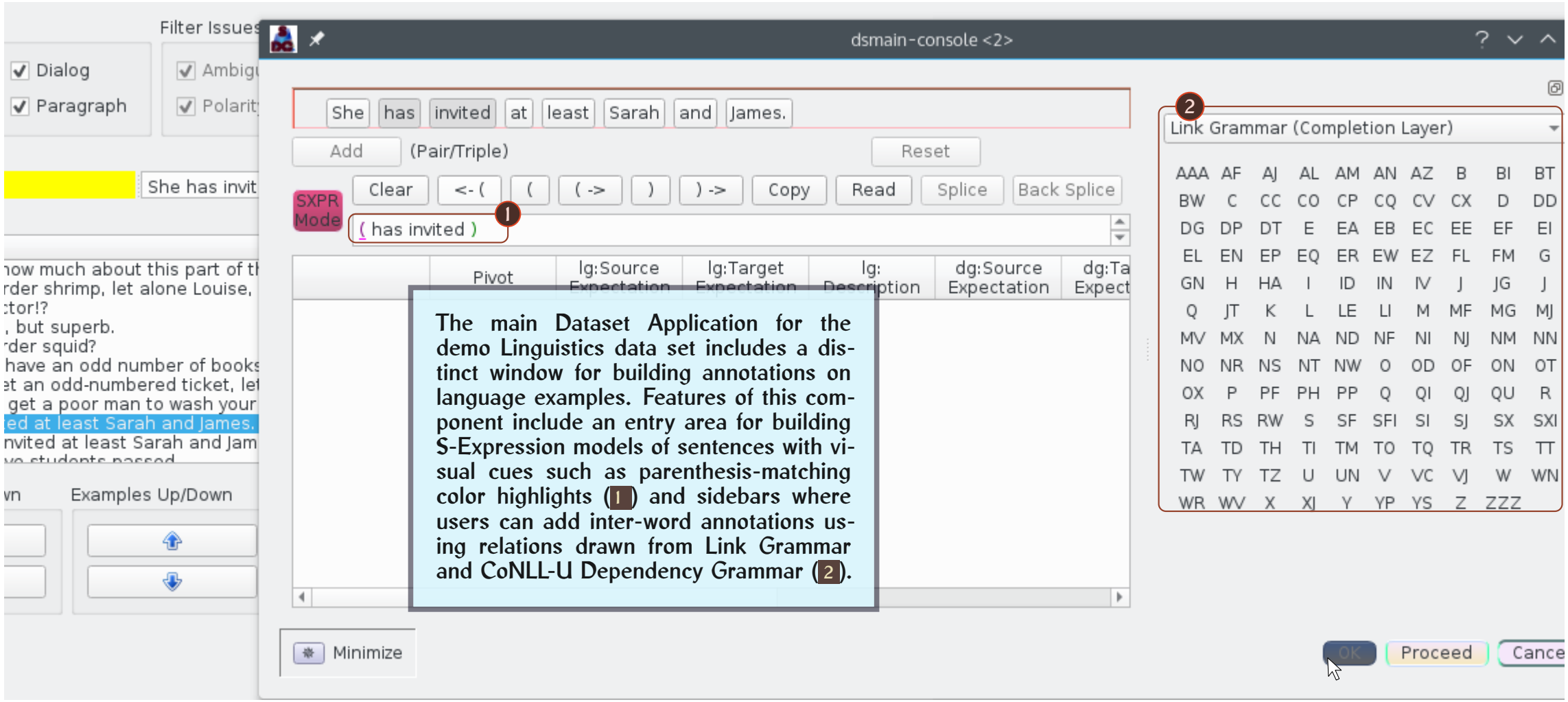
The final three screenshots show an example of how a custom-signd application can facilotat the task of building an annotated corpus from a linguistics text. The components demonstrated here enable several strategies (which can be combined) for dscribing parsing structures and the logical composition of language samples:

S-Expressions Representing linguistic units as semantic and syntactic transformations triggered by words assigned to "functional" types.

Deepndency Grammar Representing phrase structures viabinter-word syntactic relationships.

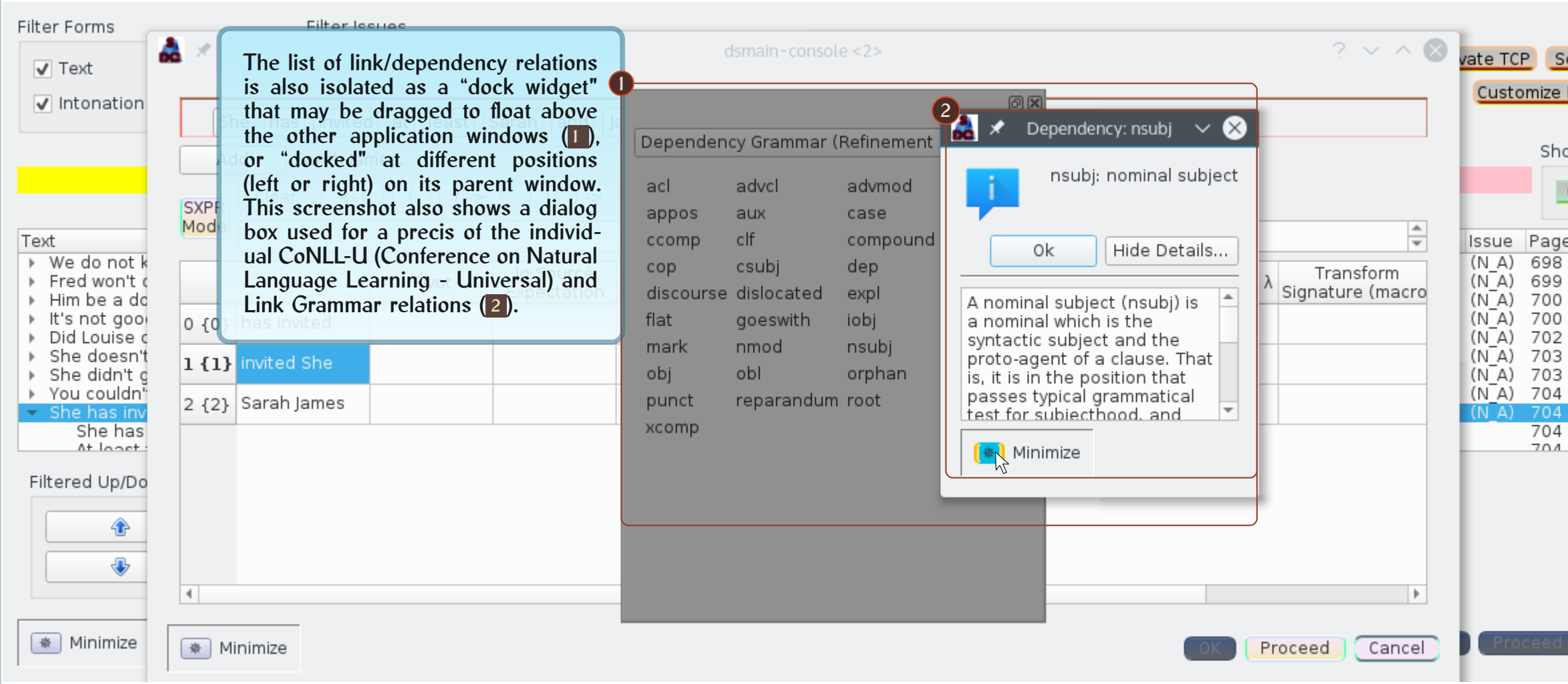
Link Grammar Representing linguistic structure via connectors internal to each word-sense. Inter-word links are activatd when each word in the pair has a connector compatible with the other word's connector. Intuitively, a connctor represents how one word's meaning or grammatic contribution can be "completed" by linking to a separate word.

Building Parsing Models

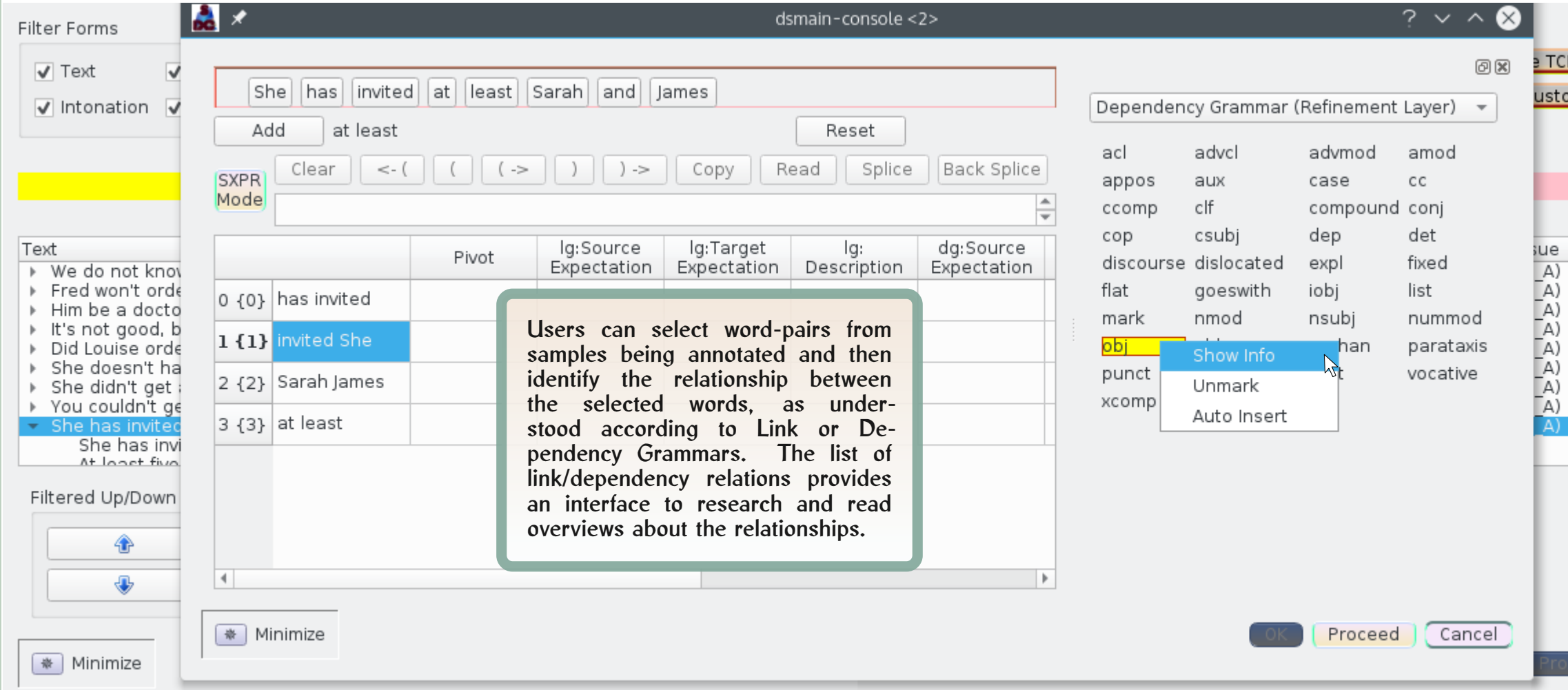


The main Dataset Application for the demo Linguistics data set includes a distinct window for building annotations on language examples. Features of this component include an entry area for building S-Expression models of sentences with visual cues such as parenthesis-matching color highlights (1) and sidebars where users can add inter-word annotations using relations drawn from Link Grammar and CoNLL-U Dependency Grammar (2).

Using Dock Widgets For Flexible Layout



Link and Dependency Grammar Annotations



Technological Components of Dataset Creator

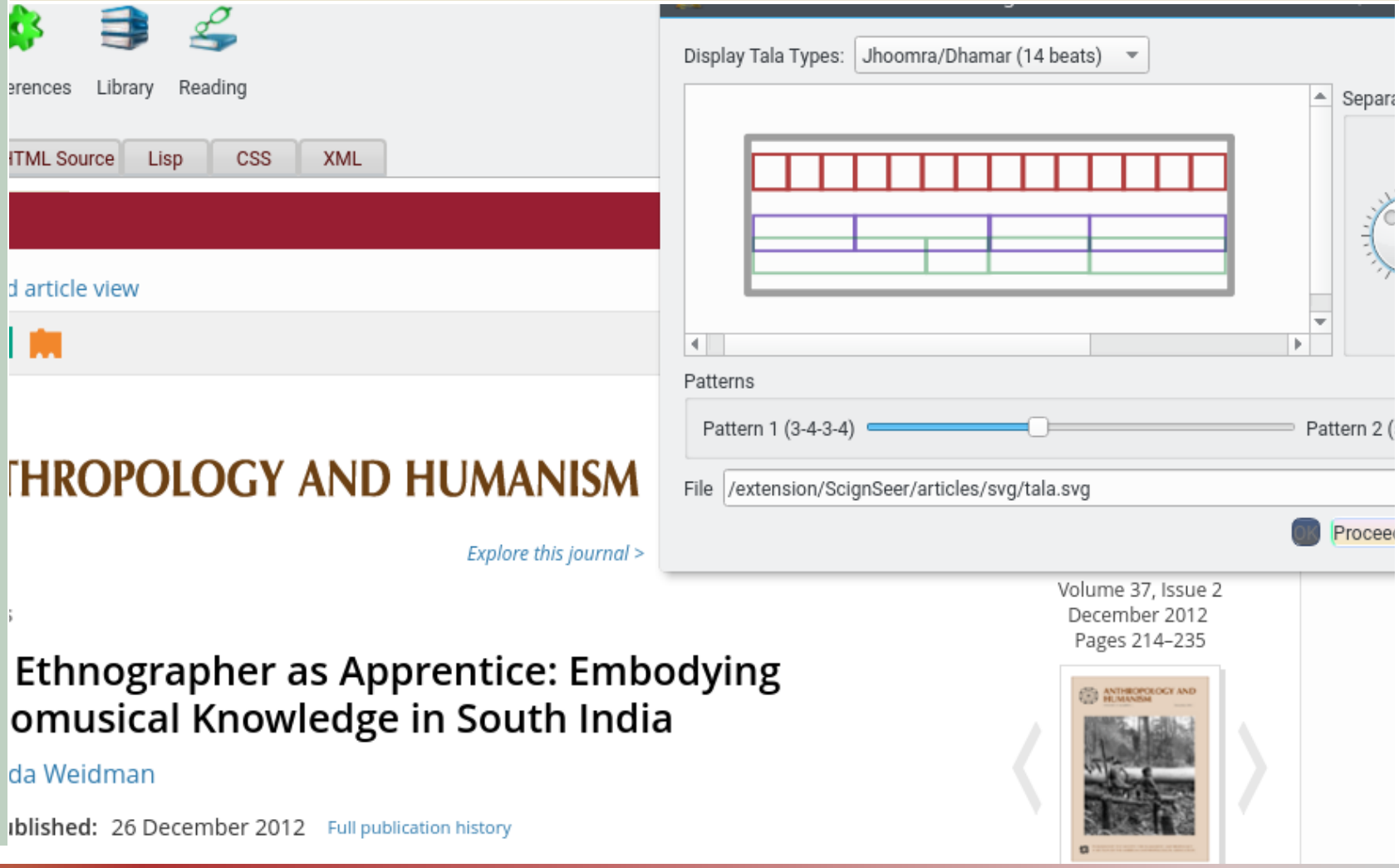
- **A3R (Application-as-a-Resource)**: A3R Applications are self-contained, citable resources and tools which can conform to modern resource documentation standards, such as the Research Object protocol. Dataset Applications can use the A3R tools and protocol to create custom custom desktop-style applications for viewing and analyzing research data, while bundling the dataset and application code into a citable Research Object.

- **HTXN (Hypergraph Text Encoding Protocol)**: HTXN is a protocol for encoding documents' character streams and document structure via "standoff annotation" (i.e., character encoding is fully separate from structural representation). HTXN supports diverse kinds of document models, including LaTeX, XML, RDF, and Concurrent/Overlapping XML extensions.

- **MOSAIC (Multiparadigm Ontologies for Scientific and Academic Publishing)**: Mosaic provides data-modeling capabilities which reflect a diversity of Information Representation paradigms, such as Hypergraphs, Conceptual Spaces, Reactive Programming, and Object-Oriented Simulation.
  - Mosaic includes the Mosaic/HTXN Semantic Document Infoset (MH-SDI) and Mosaic Plugin Framework (MPF) (see slides 27-34).

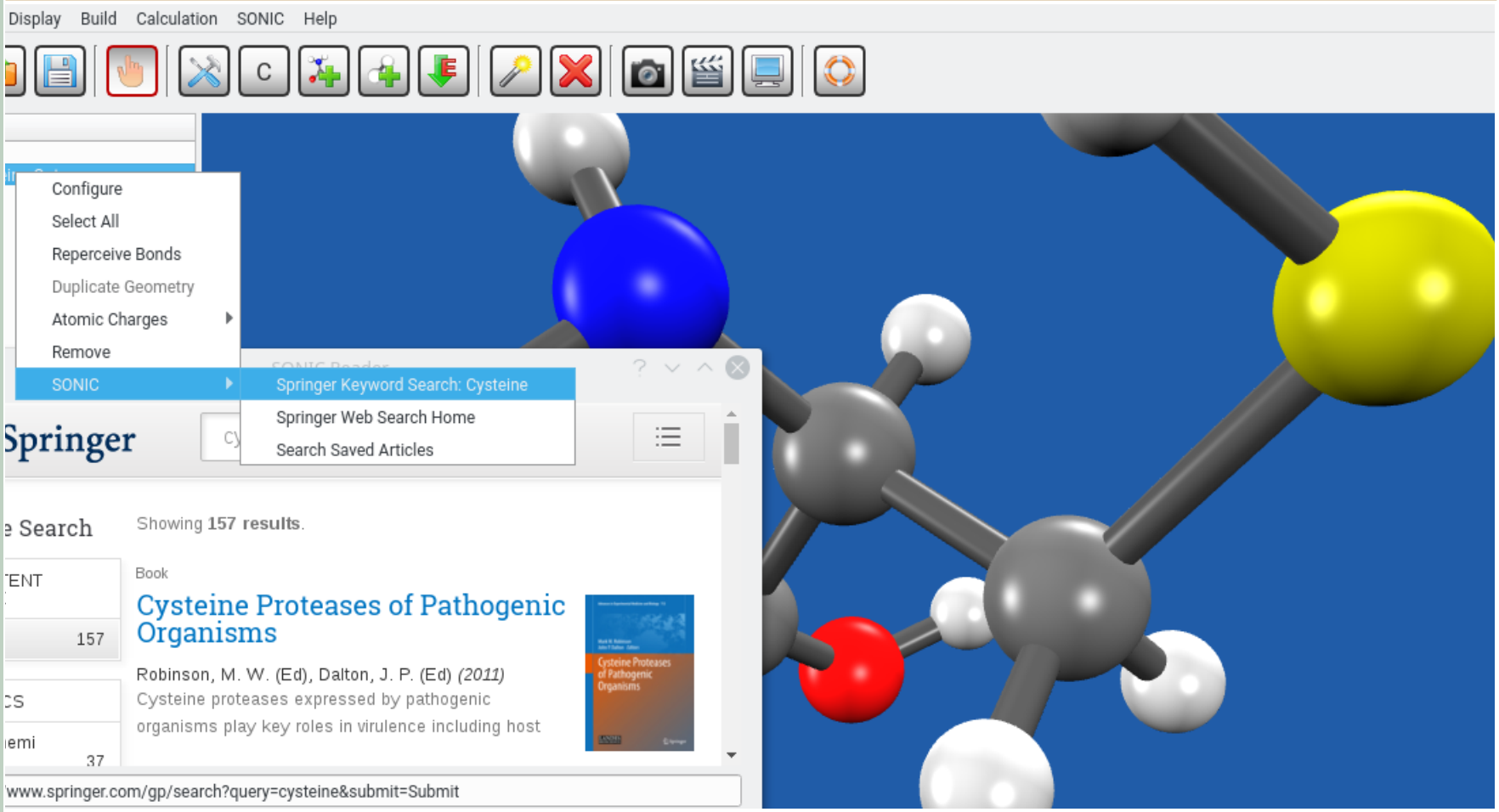
A3R Document Viewers

A3R applications may embed viewers for document formats such as e-Pub, HTML, and PDF; then supplement conventional publications with special components customized for individual manuscripts: e.g. (as in this case), a widget allowing readers to visually explore patterns in classical Indian music.



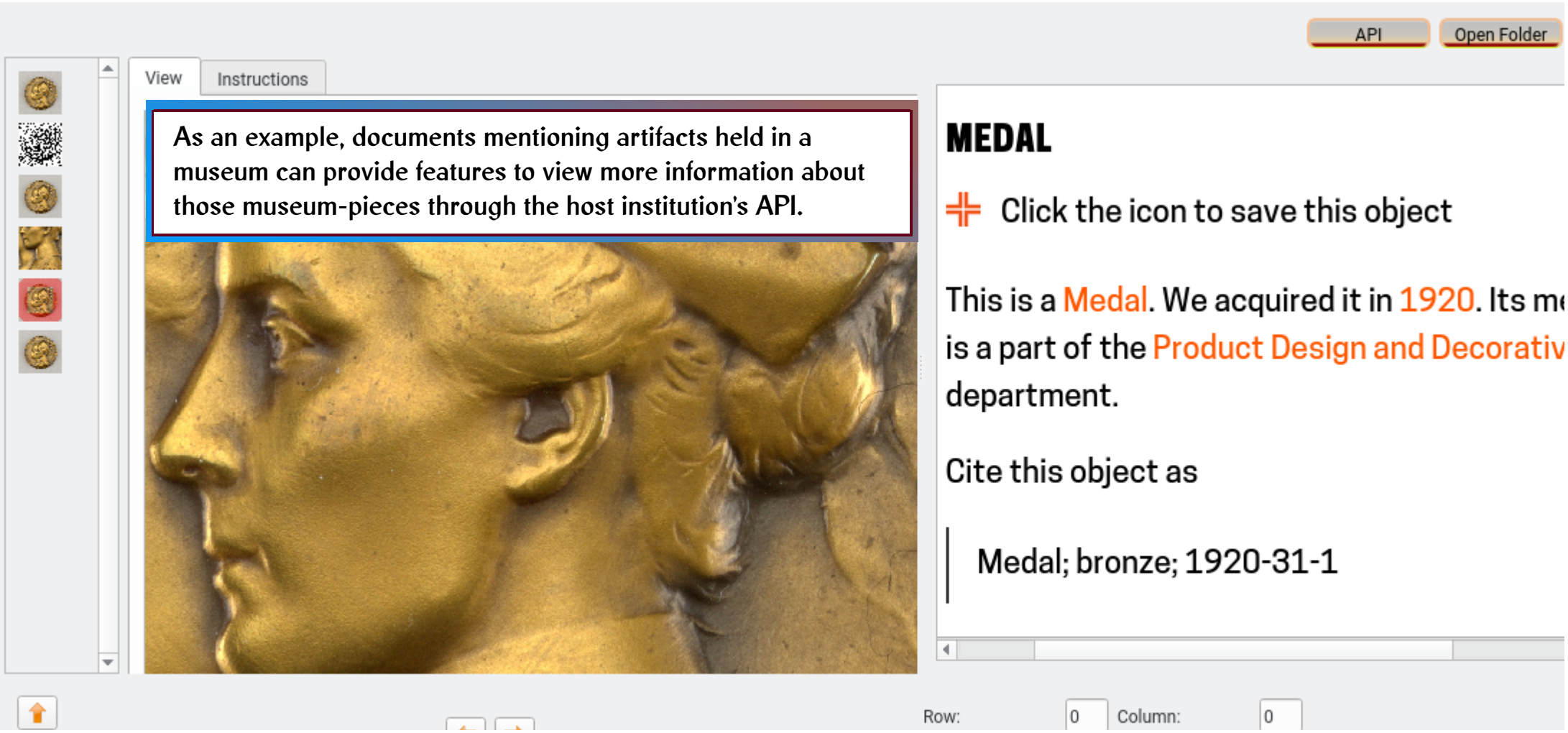
A3R Document Viewers as Embedded Components

Document Viewers may also be embedded in host applications which provide domain-specific visualization capabilities. For example, chemistry papers might be viewed within IQmol (a Qt-based program for molecular visualization and physical/chemical analysis) via an A3R document-viewer plugin.



Document Viewers Augmented With APIs

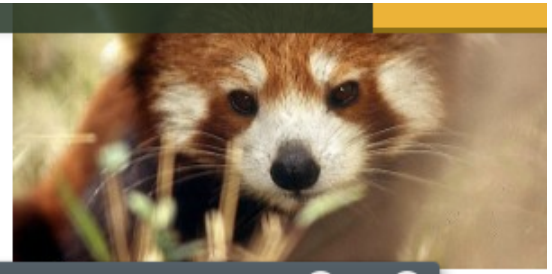
Another strategy for interactive publications is linking documents with APIs maintained by publishers, or by cultural or educational institutions.



Embedded Multimedia

Custom-built A3R document viewers can provide convenient access to multimedia content embedded in or linked to texts — including audio files, videos, and 3D graphics scenes or models.

In this case a video player is launched in a dialog box, floating above the article text. For those reading digital books or articles, videos and other multimedia content can be presented through secondary windows launched via context menus; text and multimedia may thereby be viewed side-by-side.



Ailurus fulgens styani (also known as *a. f. refulgens*): Only found in China (in the Hengduan

Behavior

Red pandas are generally solitary, but there are a couple
develop extended associations with their mothers that las
breeding season.

so they
e annua

In terms of t
tend to have
other. This n
search for th
patchily dist

They
each
das
may be
ffect the

ork.org/red_panda/about-the-red-panda/

ScignSeer Video Player

Restart Pause Play

URL file:///ext_root/videos/a.mp4

OK Proceed Cancel

Components of Mosaic

- **Mosaic/HTXN Semantic Document Infoset (MH-SDI)**: The Mosaic/HTXN Infoset is similar to an XML Infoset, embodying a machine-readable representation of documents' text, structure, and secondary resources which can be accessed according to different protocols (such as a Document Object Model). In contrast to XML, the MH-SDI supports more detailed semantic queries against document structures, such as identifying sentence boundaries and matching multimedia assets to manuscript locations.

- **Mosaic Plugin Framework (MPF)**: The Mosaic Plugin Framework is a protocol for embedding plugins or extensions within document viewers, scientific applications, and multimedia software, with the plugins inter-operating to implement multi-application networks. In particular, document viewers can launch and send data to scientific or multimedia applications so that readers can access multimedia content embedded in publications.

MOSAIC as an Alternative to Semantic Web Ontologies

Many experts have critiqued the Semantic Web for lacking conceptual rigor, adequate modeling for multi-scale information, and intrinsic representations for software requirements. To address these limitations, MOSAIC alternative Semantic Web paradigms with the following features:

Inter-Application Networking Protocol

- Interoperability is achieved by applications sharing modular and mostly autonomous code libraries that implement data models via strong typing, with (de)serialization and network/request logic implemented at the type level.
- A hypergraph-based type theory presents an overarching type-theoretic data-modeling frameworks which subsumes the type systems of most programming languages.

Multiscale, Requirements-Focused Resource Description

- Hypergraph-based Resource Framework to intrinsically support multi-scale data structures.
- Workflow-oriented “Meta-Procedure” Interface Definition framework to enforce procedural alignment among applications.
- The Mosaic networking and Resource Description protocols can be concretely implemented via the Mosaic Plugin Framework (see the following slides).

The Mosaic Plugin Framework (MPF)

MPF allows document viewers to communicate with external software, including Dataset Applications.

This slide and the next shows interop between a publication viewer (XPDF) and IQmol (a molecular visualization program). In this scenario, a student is reading practice questions for a GRE Chemistry exam. With proper supplemental data, an e-reader with MPF plugins (here XPDF) can launch a chemistry application (here IQmol) at relevant locations in the text, such as where questions involve the structure of specific molecules.

File Edit View Window Help | Mosaic

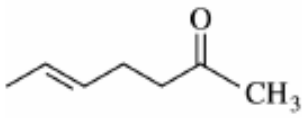
11 / 55 125% find

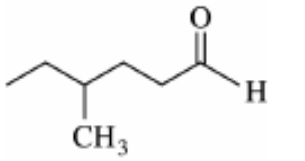
/home/nlevisrael/hypergr/ntxh/ar/

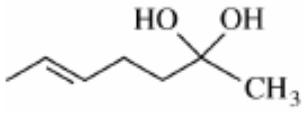
outline

- Table of Contents
  - Overview
  - Test Content
  - Preparing for the Test
  - Test-Taking Strategies
  - What Your Scores Mean
  - Taking the Practice Test
  - Scoring the Practice Test
  - Evaluating Your Perform...
  - Practice Test
  - Worksheet for Scoring th...
  - Score Conversion Table
  - Answer Sheet

1. Which of the following is the major product of the reaction shown above?

(A) 

(B) 

(C) 

(D) 

(B) Ca^{2+}
(C) Sc^{3+}
(D) Rb^{+}
(E) Sr^{2+}

4. The molecular geometry of thionyl chloride, $SOCl_2$, is best described as

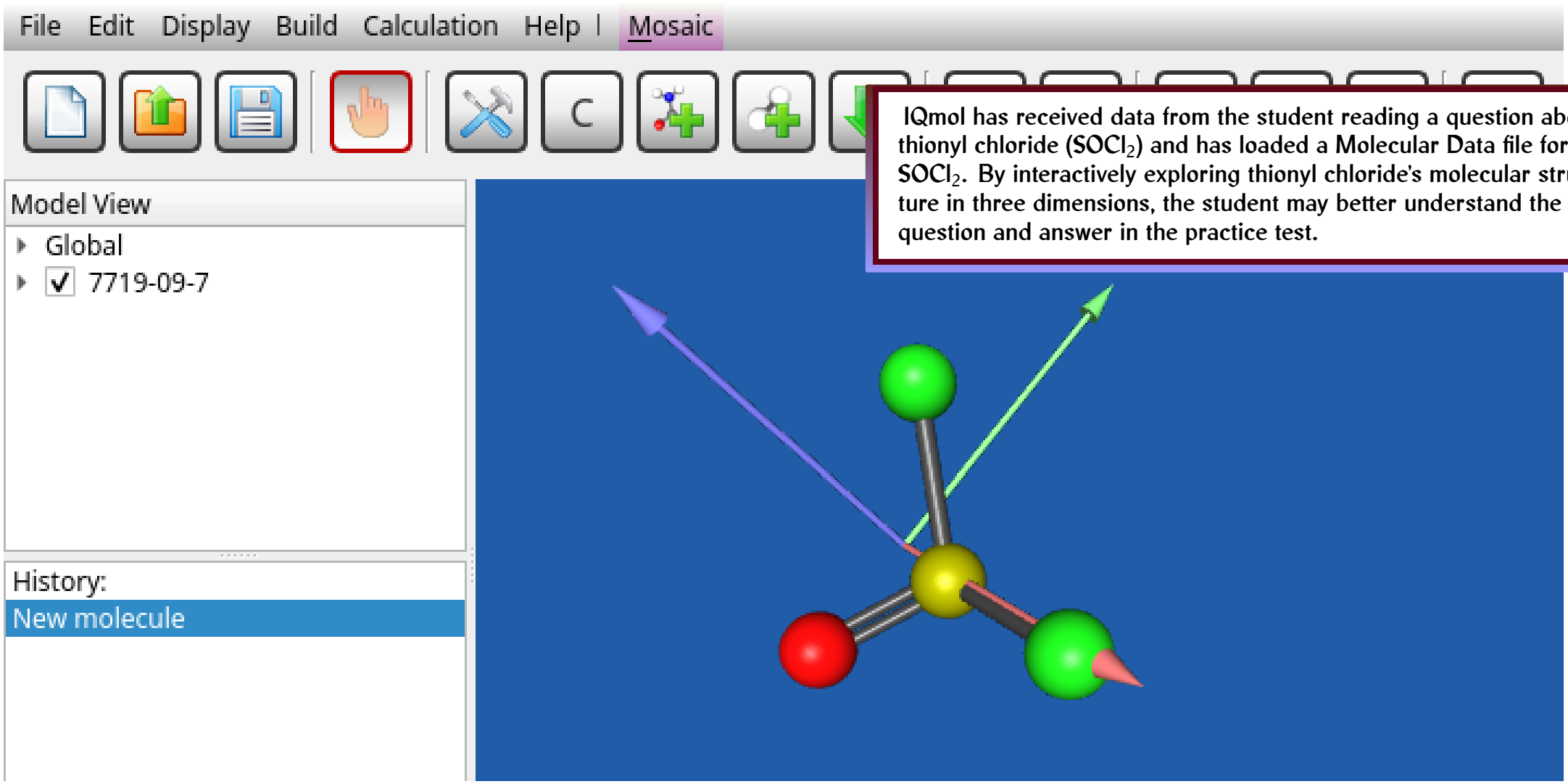
(A) trigonal planar
(B) T-shaped
(C) tetrahedral
(D) trigonal pyramidal
(E) linear

Copy "thionyl chloride"

3D thionyl chloride Viewer (launch IQmol)

Application Interop via MPF

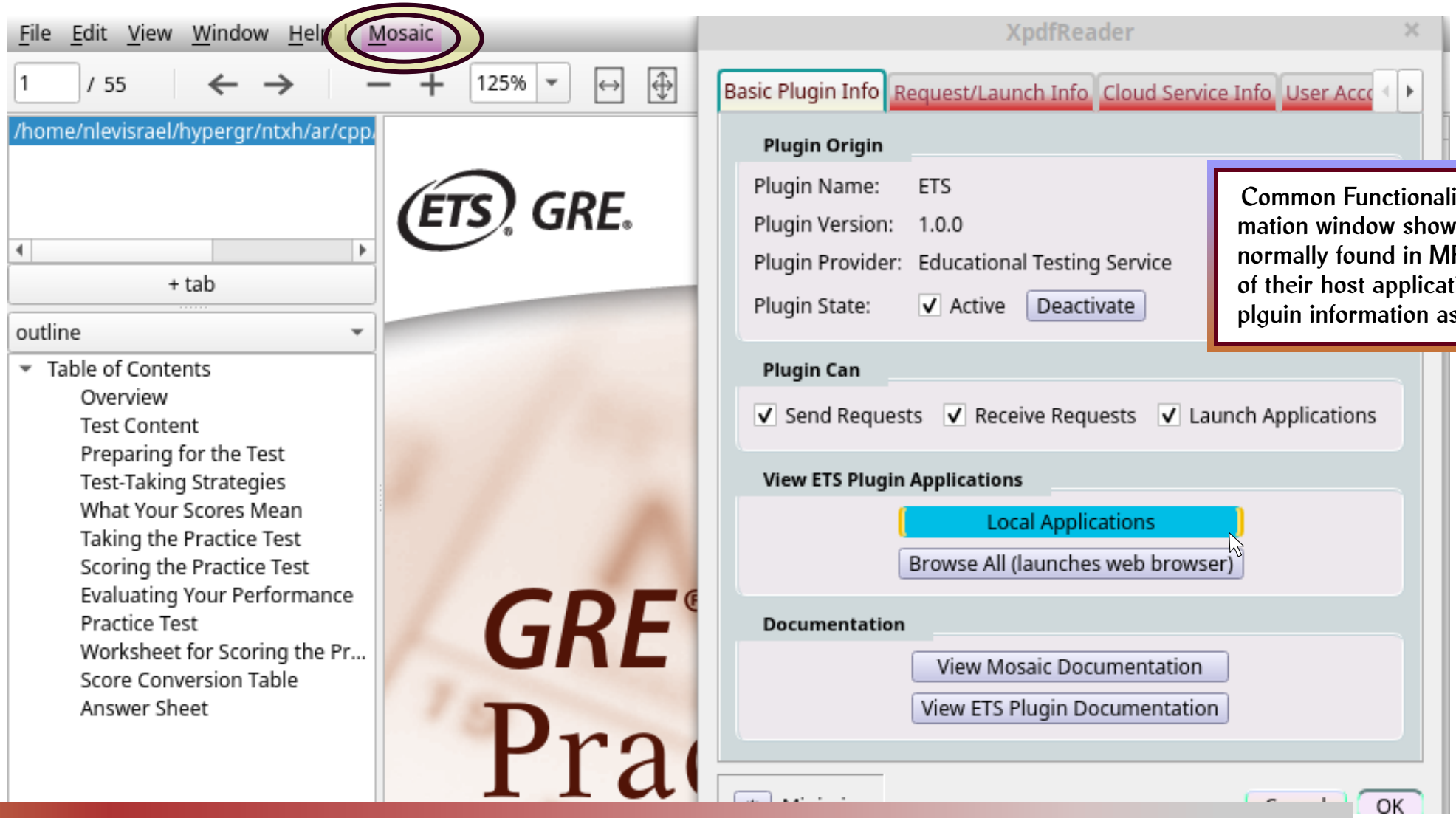
MPF allows document viewers to communicate with external software, including Dataset Applications.



IQmol has received data from the student reading a question about thionyl chloride ($SOCl_2$) and has loaded a Molecular Data file for $SOCl_2$. By interactively exploring thionyl chloride's molecular structure in three dimensions, the student may better understand the question and answer in the practice test.

Common Plugin Functionality

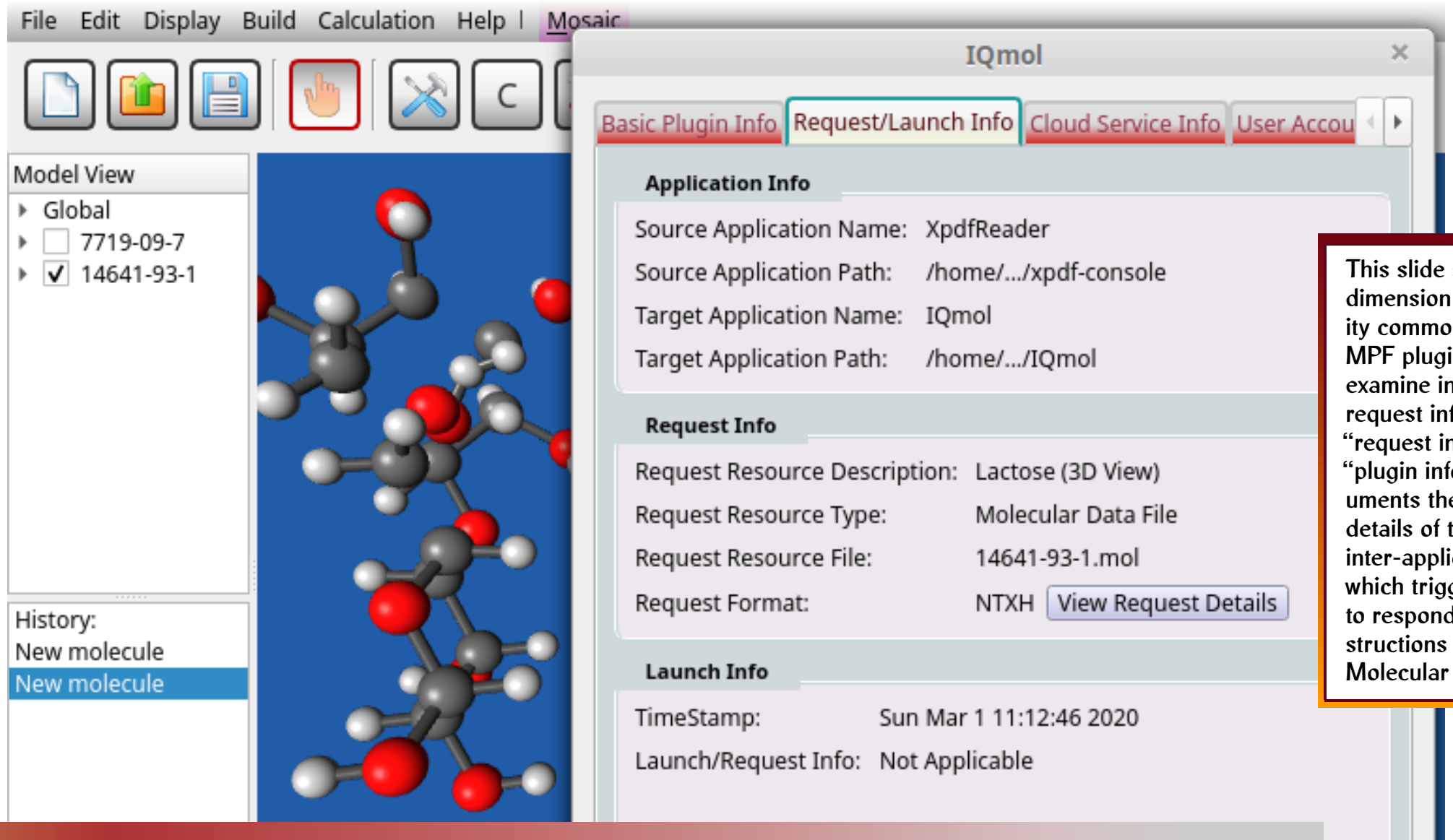
MPF Plugins have similar functionality and features in different host applications, which makes it convenient to use as readers switch among multiple applications.



Common Functionality, such as an information window showing plugin data, is normally found in MPF plugins regardless of their host application. This slide shows plguin information as seen in XPDF.

MPF Request Info

Common MPF functionality includes showing information about data sent between applications.



This slide shows another dimension of functionality common to disparate MPF plugins: the ability to examine inter-application request information. The "request info" tab on a "plugin info" dialog documents the origin and details of the most recent inter-application request which triggered the plugin to respond (in this case, instructions to load a specific Molecular Data file).

MPF Tracking User and Session Data

Plugins can store user-specific application state.

File Edit View Window Help | Mosaic

37 / 55 125% find

/home/nlevisrael/hypergr/ntxh/ar/cpp/pract

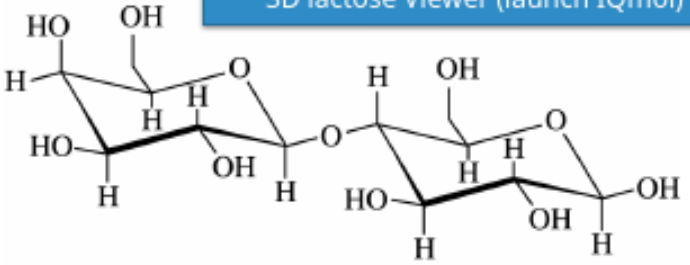
+ tab

outline

- Table of Contents
  - Overview
  - Test Content
  - Preparing for the Test
  - Test-Taking Strategies
  - What Your Scores Mean
  - Taking the Practice Test
  - Scoring the Practice Test
  - Evaluating Your Performance
  - Practice Test
  - Worksheet for Scoring the Practice...
  - Score Conversion Table
  - Answer Sheet

Copy "lactose"

3D lactose Viewer (launch IQmol)



95. Which of the following is NOT true about the disaccharide lactose shown above?

(A) Lactose is a reducing sugar.
(B) Lactose undergoes mutarotation.
(C) Lactose is optically active.
(D) Lactose can be hydrolyzed to monosaccharides with H_2O/H_2SO_4.
(E) Lactose has a 1,1′-α- glycosidic linkage.

AAK ALL

I II

97. A peptide digest yield
listed above. The thre
using capillary electro
at which each peptide
charge. Which of the
order, from first to las
reach the detector? (A
G = glycine; K = lysi

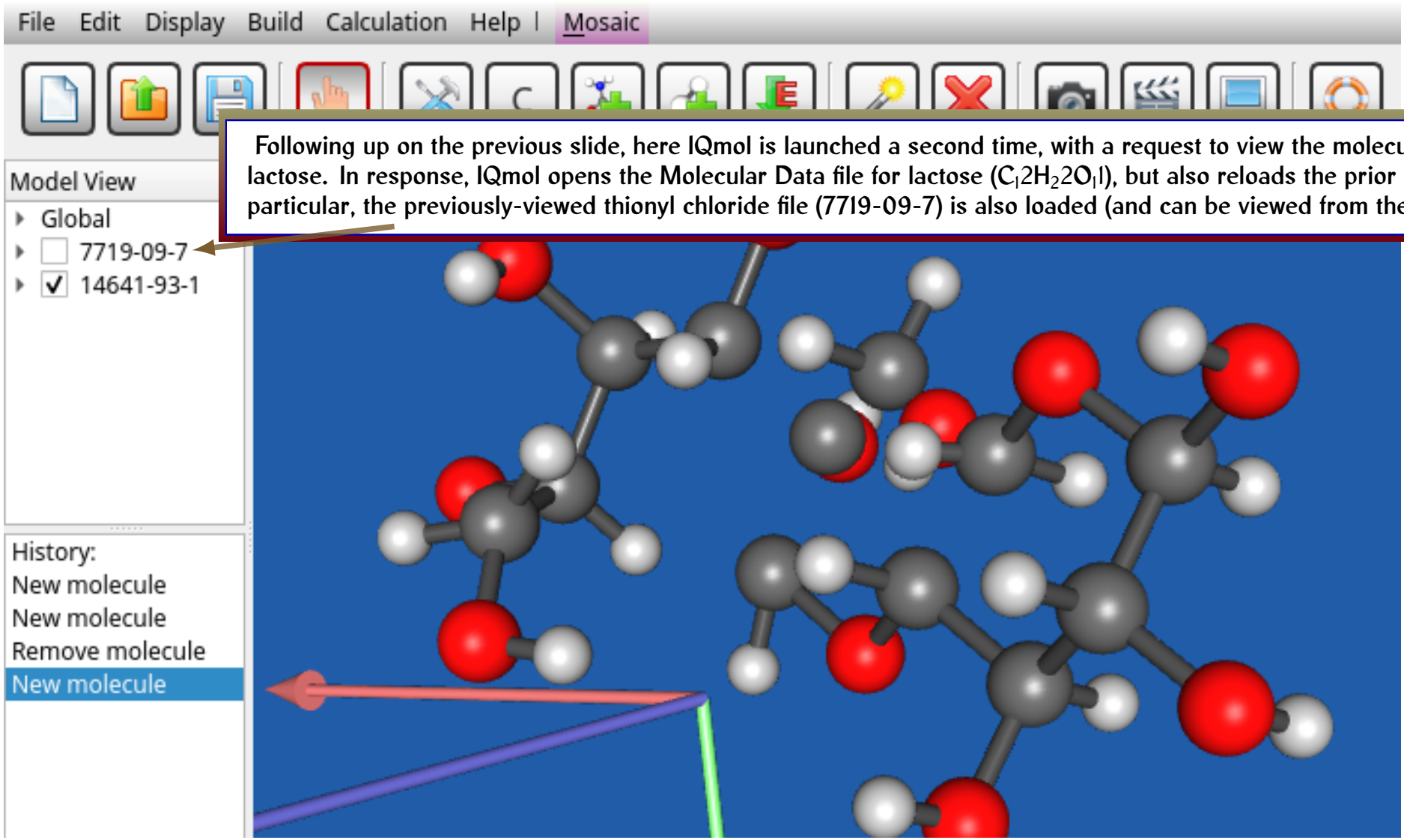
(A) I, II, III
(B) I, III, II
(C) II, I, III
(D) II, III, I
(E) III, II, I

98. In fluorescence spectroscopy, (Φ_f) is best defined as the

Plugins can remember previous interactions between applications and can send detailed information packages — inter-application networking is not limited to sending single multimedia files. In this slide a student is seen launching IQmol from a later question in the GRE practice test ...

Reloading User Sessions

After being launched a second time, the MPF plugin can reload the prior application state.



Following up on the previous slide, here IQmol is launched a second time, with a request to view the molecular structure of lactose. In response, IQmol opens the Molecular Data file for lactose ($C_{12}H_{22}O_{11}$), but also reloads the prior session — in particular, the previously-viewed thionyl chloride file (7719-09-7) is also loaded (and can be viewed from the side panel).

Thank You!

Please contact Linguistic Technology Systems for more information about dsC and/or other Software Development and Software Language Engineering Solutions: (917) 817-2184.