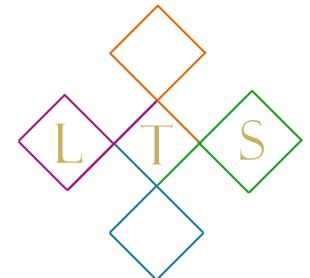
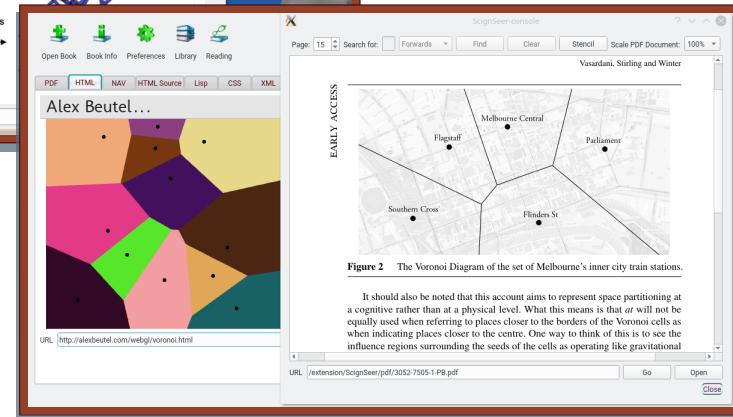
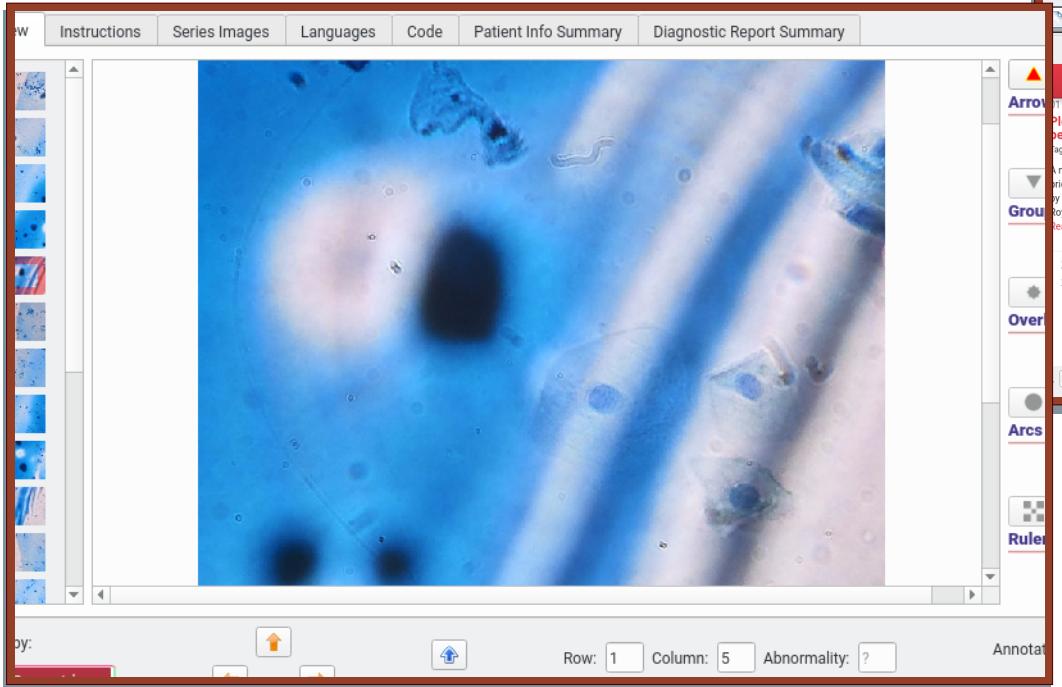


# Dataset Creator

## ("dsC")



Linguistic Technology Systems (LTS)  
Amy Neustein, Ph.D., Founder and CEO  
[amy.neustein@verizon.net](mailto:amy.neustein@verizon.net)  
(917) 817-2184  
March 6, 2020

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/desktop 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.

# 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, such as:

**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

[Customize Build](#)[Activate TCP](#)[Screenshot](#)

Main Flow Temperature Oxygen

Index	Flow	Time With / Average	Time Against / Delta	Temperature C° / K°	Oxygen (calculated)
1	0.561	0.000219893	0.000220329	49.60	
2	1.17	0.000219764	0.000220614	49.70	
		0.000220189	8.49999e-7	322.15	93
	%	0.106536		67.3623	1
	#	159		322	394
3	5.133	0.000218866	0.000221751	49.70	
4	10.89	0.000218223	0.00022191	48.90	
5			0.000218854	49.50	
6			0.000219006	49.60	

One rationale for using nested rows is for displaying supplemental information, such as data values' rank (3) and percentage (2) (on the scale of the least to greatest value) relative to all other values for each statistical parameter.

Using a "tree widget" (a two-layer spreadsheet), instead of a conventional spreadsheet, allows the Dataset Application to distinguish primary values (those measured directly by physical devices and experimental equipment) from intermediate values calculated via algorithms.

Sample Up/Down

Peer Up/Down

First



Peer First



DOUBLE

Graphics

- 2D 25x25
- 2D 12x12
- 2D 3x3
- 2D 37x75
- 3D 25x25
- 3D 12x12
- 3D 3x3
- 3D 37x75

# Interacting with the Main Window

Customize Build Activate TCP Screenshot

The screenshot shows a dataset application window with a table of data. The tabs at the top are Main, Flow, Temperature (selected), and Oxygen. The table has columns: Index, Flow, Time With / Average, Time Against / Delta, Temperature C° / K°, and Oxygen (calculated). A context menu is open over the table, with items: About/ Show in Document (may require XPDF), Copy Column to Clipboard (values) (selected), and Copy Column to Clipboard (ranks). Numbered callouts point to: (1) the Temperature tab, (2) the context menu, (3) the Peer Up/Down buttons, and (4) the Sample Up/Down buttons. A callout box states: "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." Another callout box states: "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)."

Index	Flow	Time With / Average	Time Against / Delta	Temperature C° / K°	Oxygen (calculated)
33	0.589	0.00022861	0.000229046	5.40	
		0.000228828	4.35997e-7	278.15	80
%	0.0531...			7.25373	0
#	111			1	34
34	1.098	0.000228924	0.000229746	5.40	
39	4.988	0.000228814	0.000231814	5.40	
35	5.044	0.000227894	0.000230985	5.40	
37	0.554	0.000229983	0.00023039	5.50	
38	1.057	0.000229819	0.000230657	5.50	
31	5.057	0.000229433	0.000232403	5.50	
30	1.108	0.000230476	0.000231223	5.70	
29	0.484	0.000230511	0.000230934	5.80	

Sample Up/Down

Peer Up/Down

First

Peer First

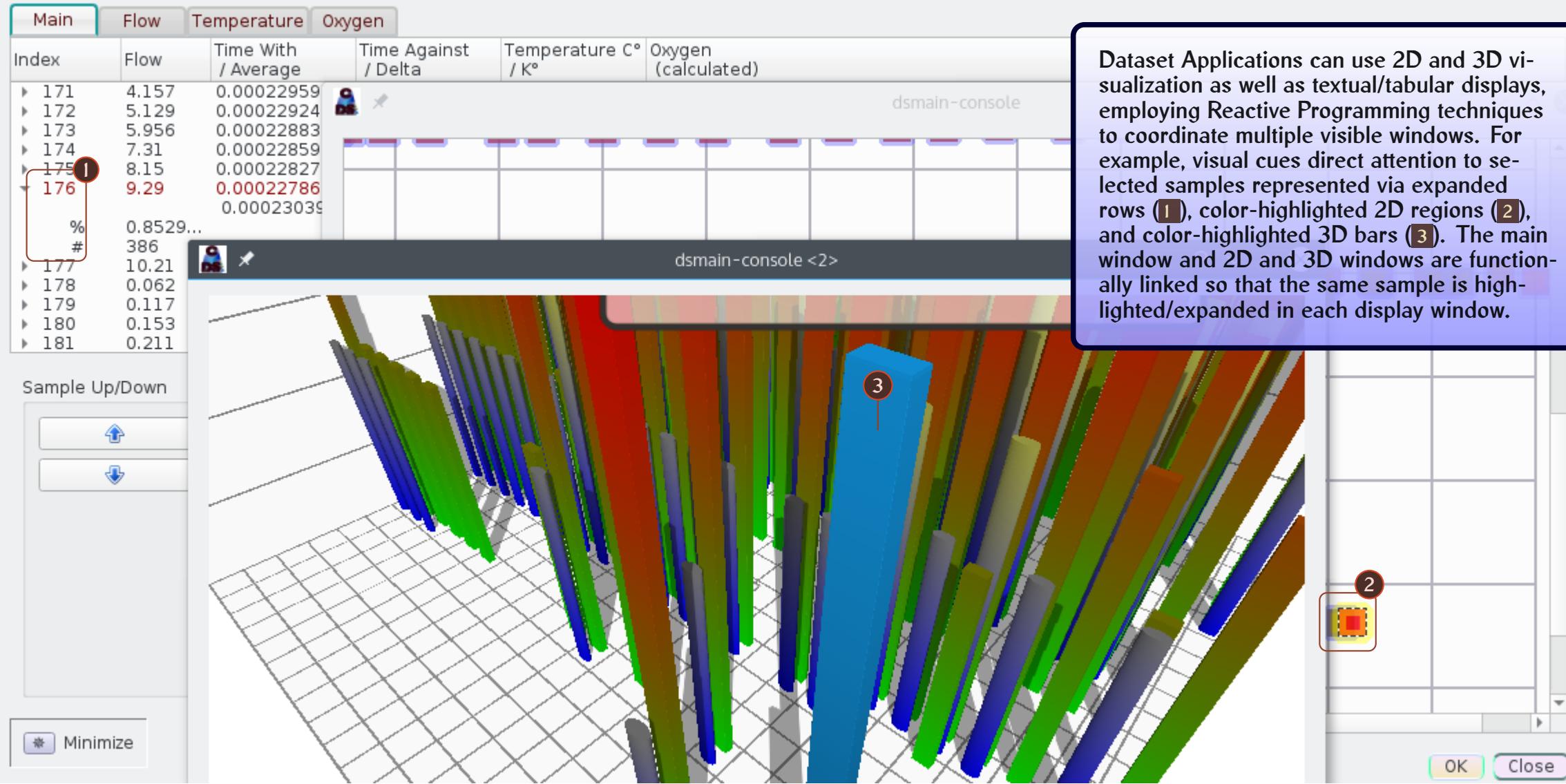
2D 25x25 2D 12x12 2D 3x3 2D 37x75

3D 37x75

OK Proceed Close

# Coordinated Data Visualization

Customize Build Activate TCP Screenshot



# Interacting with the Visuals

Customize Build Activate TCP Screenshot

Main Flow Temperature Oxygen

Index	Flow	Time With / Average	Time Against / Delta	Temperature C° / K°	Oxygen (calculated)
171	4.157	0.00022959			
172	5.129	0.00022924			
173	5.956	0.00022883			
174	7.31	0.00022859			
175	8.15	0.00022827			
176	9.29	0.00022786			
		0.00023039			
%	0.8529...				
#	386				
177	10.21	0.00022762			
178	0.062	0.00022844			
179	0.117	0.00022852			
180	0.153	0.00022852			
181	0.211	0.00022905			

Sample Up/Down Peer Up/Dow

Minimize Minimize

dsmain-console

0.011 0.917 1.824 2.730 3.637 4.543 5.450 6.3

5.400

11.54

17.68

23.82

29.96

0 1 2 3 4 5 6 7

1 2 3

OK Close

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 shrinking all others (3).

Scroll to Top Left  
Contract Nearby Items (1 cell)  
Contract Nearby Items (2x2 cells)  
Contract Nearby Items (2xAll cells)  
Contract Nearby Items (4x4 cells)  
Uncontract (All cells)  
Highlight Oxygen = 93  
Highlight Oxygen = 90  
Highlight Oxygen = 87  
Highlight Oxygen = 80  
Unhighlight Oxygen

# Research Protocols and 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 subsequent 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

[Customize Build](#)[Activate TCP](#)[Screenshot](#)

Main	Flow	Temperature	Oxygen
Index	Flow	Time With / Average	Time / Delta
▶ 33	0.589	0.00022861	0.00022861
▼ 34	1.098	0.000228924	0.000228924
		0.000229335	8.220000000000001
%	0.0999...		
#	154		
▶ 39	4.988		
▶ 35	5.044		
▶ 37	0.554		
▶ 38	1.057		
▶ 31	5.057		
▶ 30	1.108		
▶ 29	0.481		

Sample Up/Down



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.

Flow of Oxygenated Air

Click 'Show Details' for a summary or 'More' for PDF/Original Article links.

More (PDF) ... Cancel Hide Details...

The Flow measurements calculate the flow of oxygenated air (as needed for Continuous Positive Airway Pressure (CPAP) devices) given inputs of ambient temperature and sound time travel. The third (nested) row beneath the Flow value shows each sample's Flow 'rank' (where lower ranks mean that a sample has less Flow; the rank #1 is the sample with least flow). The second nested value shows each sample's flow measurement as a fraction of the maximum measurement

Minimize

OK

Proceed

Close

# Embedding XPDF

[Customize Build](#)[Activate TCP](#)[Screenshot](#)

Each data set can be linked back to an original article or other publication reporting on the data set and experimental results. Different parts of the data set can be linked to textual anchors in the publication.

2 of 21

WILEY Expert Systems

because we know that air is a relatively fixed mixture of gases, primarily consisting of nitrogen, oxygen, argon, and carbon dioxide, that in varying amounts of water vapour or humidity. The speed of sound in air is approximately 343 m/s at room temperature (20 °C or 70 °F). This is primarily a function of temperature; the only other factor that has a significant influence is the amount of humidity in the air. However, humidity has only a slight influence; an increase in relative humidity by only a small amount, say 0.5%, we can conclude that the speed of sound travels slower at higher altitudes. This is because the temperature and relative humidity are lower and not the air pressure is lower at higher altitudes. The speed of sound goes up concurrently.

In this example, after viewing a short description of a particular data field inside the Dataset Application, researchers have the option of studying that parameter further by reading at the location in the original publication where the field is introduced or described (1). The XPDF viewer is compiled as an embedded application within the main Dataset Application and can itself be customized for each data set.

PROPOSAL ASPECTS

# Testing and Fine-Tuning Dataset Applications

## 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.

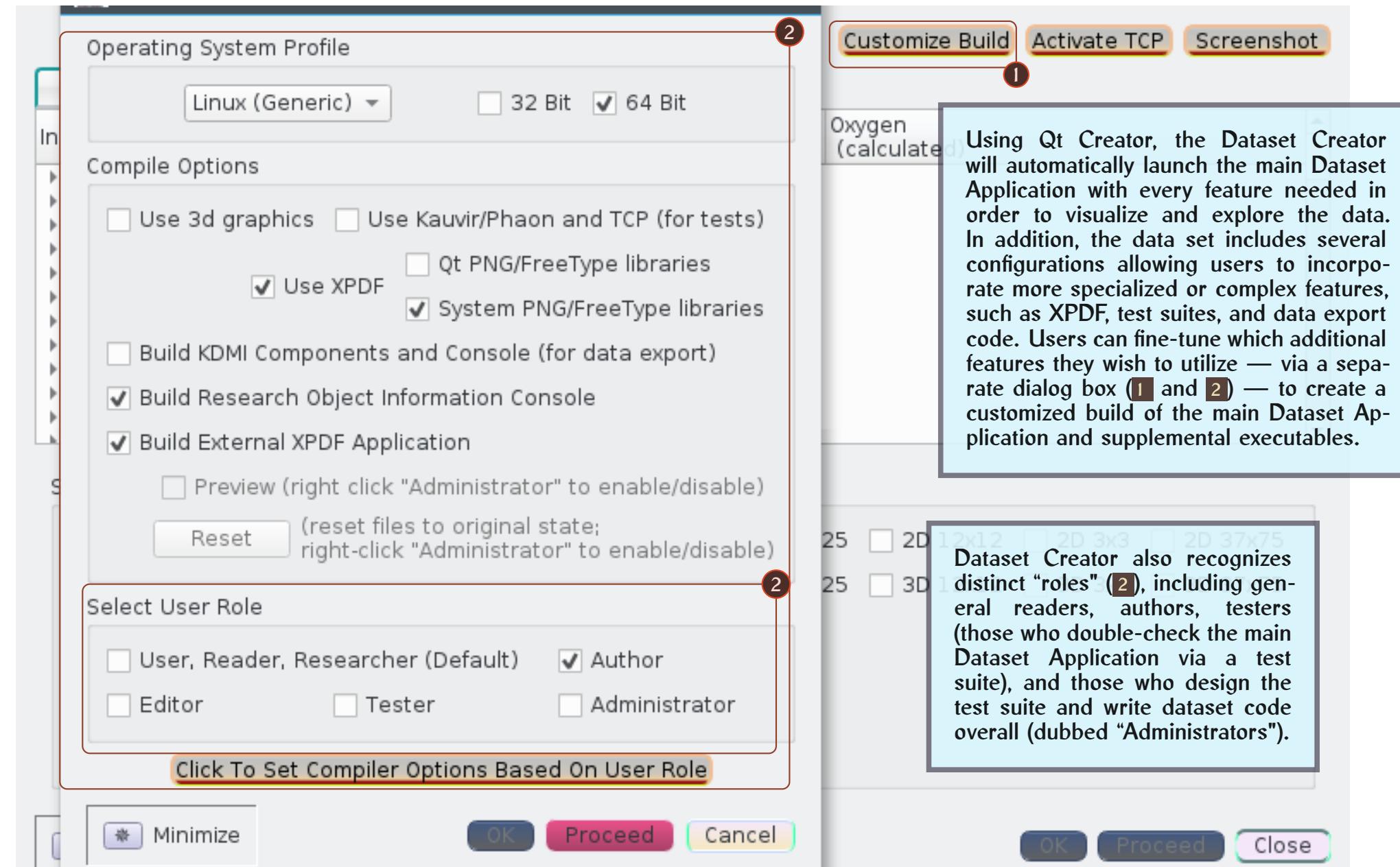
**Data Export** Tools for reusing data in other projects.

**External Libraries** Some features — such as XPDF and 3D graphics — require supplemental libraries that cannot be published with the data set in source code form. Advanced users can select which of these libraries to incorporate into their version of the Dataset Application.

**Scripting** Data sets can compile 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



Operating System Profile

Linux (Generic)   64 Bit

Compile Options

Use 3d graphics  Use Kauvir/Phaon and TCP (for tests)  
 Use XPDF  Qt PNG/FreeType libraries  
 Build KDMI Components and Console (for data export)  
 Build Research Object Information Console  
 Build External XPDF Application  
 Preview (right click "Administrator" to enable/disable)

(reset files to original state;  
right-click "Administrator" to enable/disable)

Select User Role

User, Reader, Researcher (Default)  Author  
 Editor  Tester  Administrator

Click To Set Compiler Options Based On User Role

Customize Build  Activate TCP  Screenshot

Oxygen (calculated)

25  2D 12x12 2D 3x3 2D 37x75  
25  3D 12x12 2D 3x3 2D 37x75

OK Proceed Close

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.

Dataset Creator also recognizes distinct "roles" (2), including general readers, authors, testers (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).

Customize Build (1) | Activate TCP (1) | Screenshot (1)

Test Returned (3)

Test Copy Temperature Ranks: Pass or Fail?

Pass | Fail | Hide Details...

Note: For tests which involve values copied to the system clipboard, you can use the text area below as a scratch pad to examine the clipboard contents.

Clipboard Content:

- 318
- 322
- 323
- 284
- 317

OK | Proceed | Close

Copy Temperature Ranks: This test should result in the Temperature ranks (sorted by index) being copied to the system clipboard, which can be verified by pasting the clipboard into a blank file and comparing the lines (there should be one sample per line) to the Temperature column as viewed in the tree table dialog.

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

OK

# 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 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 L<sup>A</sup>T<sub>E</sub>X, 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, and Object-Oriented Simulation.
  - Mosaic includes the Mosaic/HTXN Semantic Document InfoSet (MH-SDI) and Mosaic Plugin Framework (MPF) (see slides 19-26).

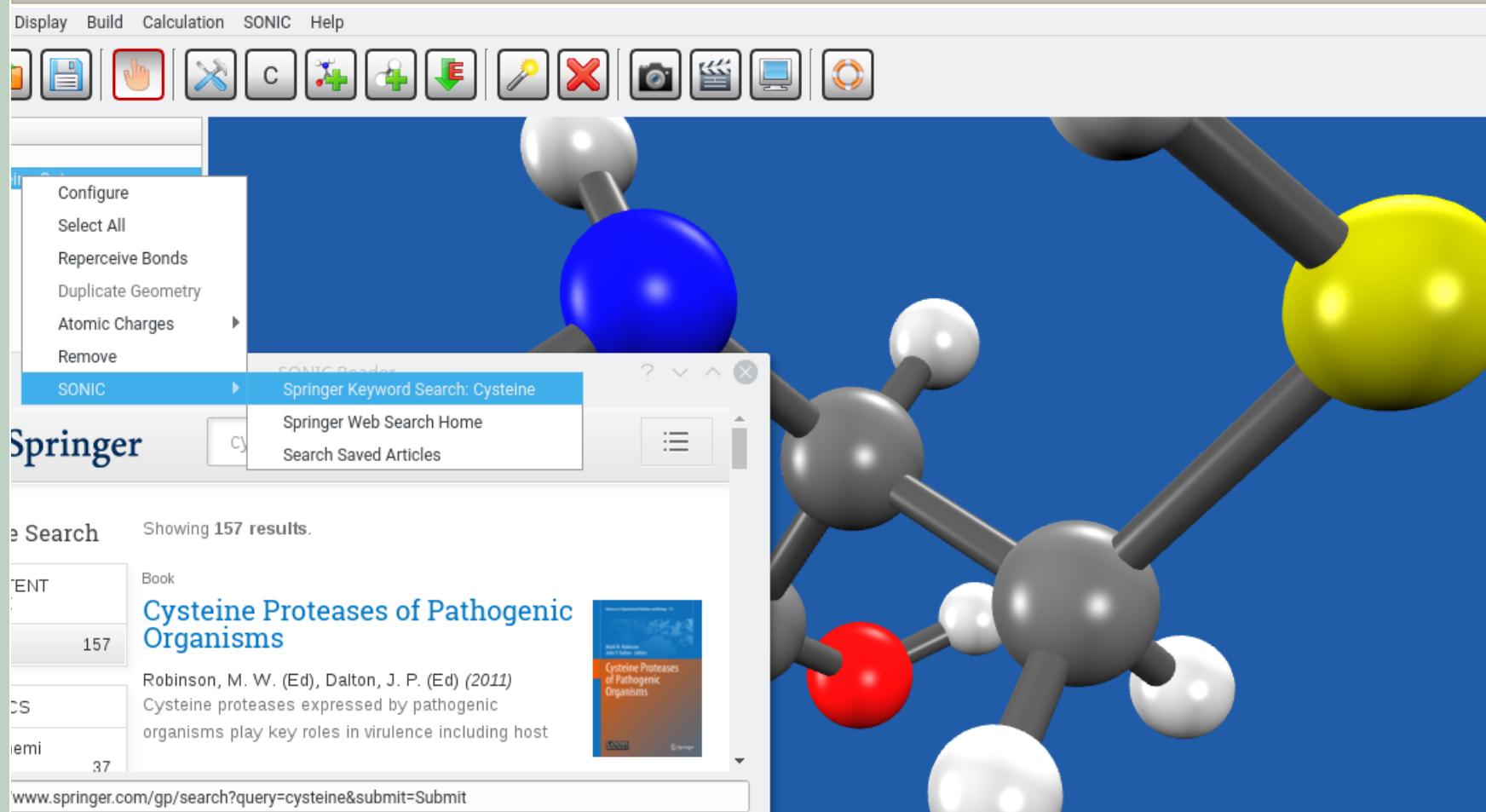
# A3R Document Viewers

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.

The image shows a screenshot of a journal website for 'ANTHROPOLOGY AND HUMANISM'. The top navigation bar includes icons for settings, library, and reading, and tabs for HTML Source, Lisp, CSS, and XML. The main content area features the journal's title 'ANTHROPOLOGY AND HUMANISM' and a link to 'Explore this journal >'. Below this is an article title 'Ethnographer as Apprentice: Embodying musical Knowledge in South India' by 'Laura Weidman' published on '26 December 2012' with a link to 'Full publication history'. A central feature is a document viewer window titled 'Display Tala Types: Jhoomra/Dhamar (14 beats)'. This window contains a grid of colored rectangles (red, purple, green) representing musical patterns. Below the grid, a slider is labeled 'Pattern 1 (3-4-3-4)' on the left and 'Pattern 2 (4-4-4-4)' on the right. The file path '/extension/ScignSeer/articles/svg/tala.svg' is shown. At the bottom of the viewer window are 'OK' and 'Proceed' buttons. To the right of the viewer, journal metadata is displayed: 'Volume 37, Issue 2', 'December 2012', and 'Pages 214-235'. A thumbnail of the journal cover is shown, which features a person in a traditional setting.

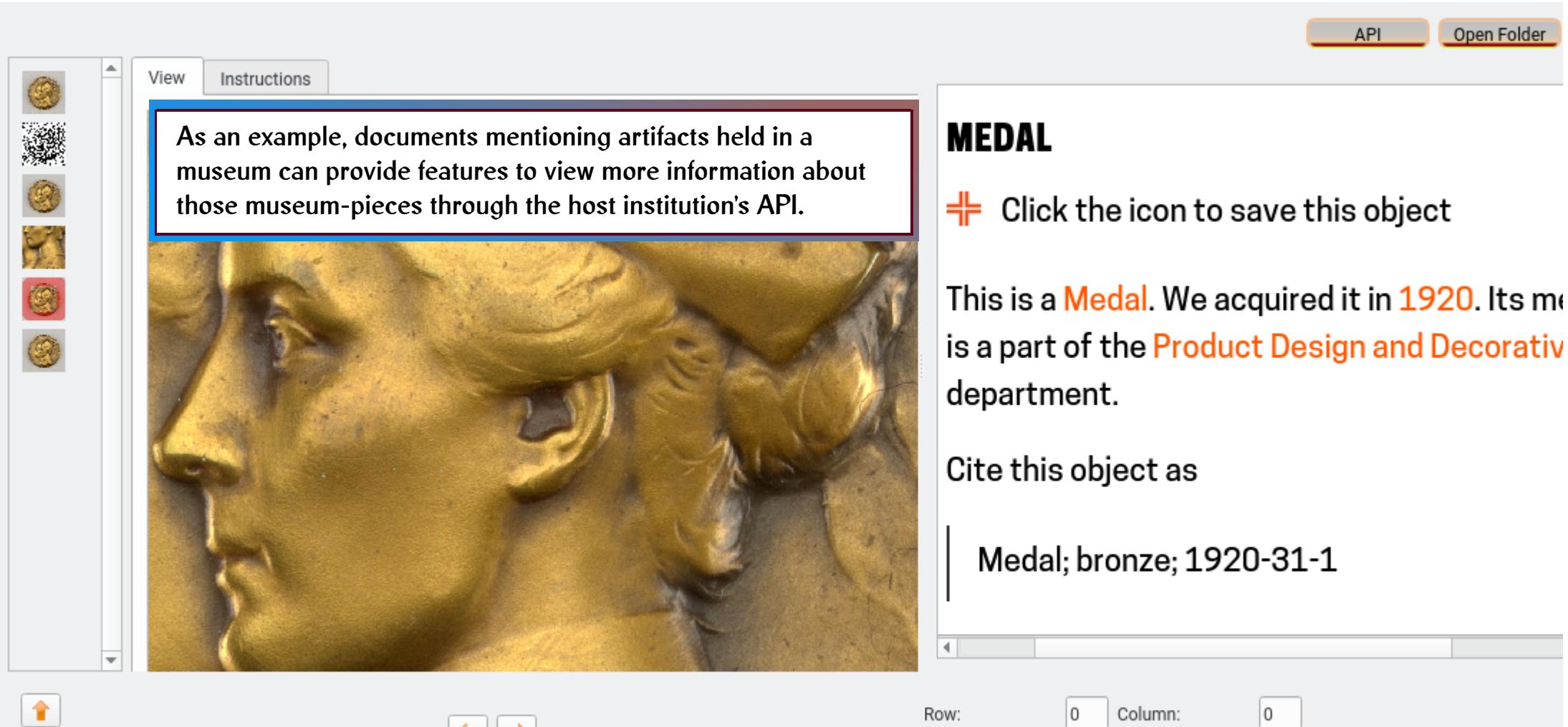
# A3R Document Viewers as Embedded Components

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.



The screenshot shows a digital document viewer interface. At the top, there are buttons for 'View' and 'Instructions'. On the right, there are 'API' and 'Open Folder' buttons. On the far left, there is a vertical sidebar with several small thumbnail images of different artifacts. The main content area features a large image of a bronze medal in profile, showing a person's head. Overlaid on this image is a red-bordered box containing the text: 'As an example, documents mentioning artifacts held in a museum can provide features to view more information about those museum-pieces through the host institution's API.' To the right of the image, the word 'MEDAL' is displayed in large, bold, black capital letters. Below it, there is a red icon of a plus sign inside a circle, followed by the text 'Click the icon to save this object'. Further down, the text 'This is a **MEDAL**. We acquired it in **1920**. Its me is a part of the **Product Design and Decorativ** department.' is visible. At the bottom, there is a section for citation with the text 'Cite this object as' and a reference 'Medal; bronze; 1920-31-1'. At the very bottom of the interface, there are buttons for 'Row:', 'Column:', and '0'.

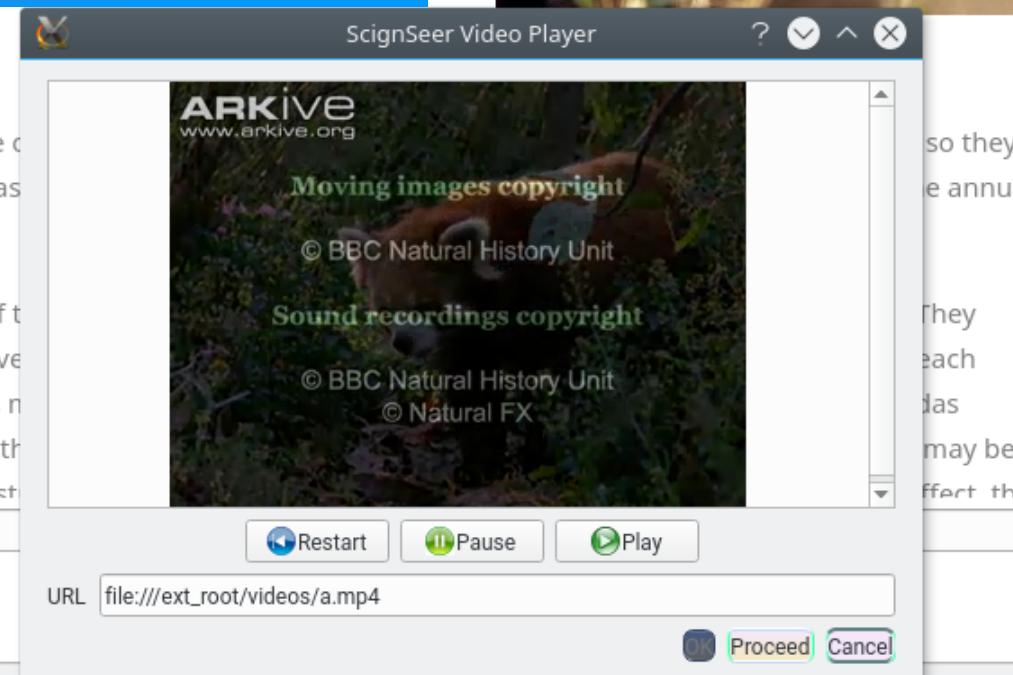
# Embedded Multimedia

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.

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

Mo  
My  
The  
ab

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.



## Behavior

Red pandas are generally solitary, but there are a couple of cases where they develop extended associations with their mothers that last through the breeding season.



In terms of territoriality, red pandas tend to have overlapping home ranges with other. This may be because they search for the same food items, which are often patchily distributed.

[arkive.org/red\\_panda/about-the-red-panda/](http://arkive.org/red_panda/about-the-red-panda/)

## 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. Documents composed with HTXN embed PDF page coordinates for semantically significant text locations (such as sentence boundaries), allowing PDF viewer plugins to provide supplemental features (such as dataset links or "review questions" joined to individual sentences, paragraphs, or figure illustrations).
- ◆ **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 encompasses 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.

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)

(E)

4. The molecular geometry of thionyl chloride,  $\text{SOCl}_2$ , 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.

File Edit Display Build Calculation Help | Mosaic



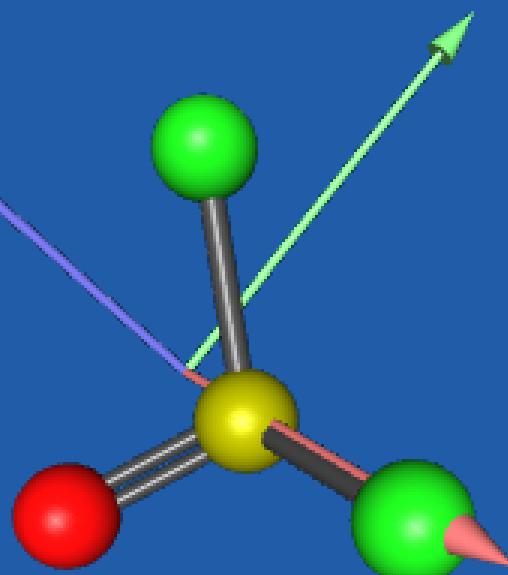
Model View

- ▶ Global
- ▶  7719-09-7

History:

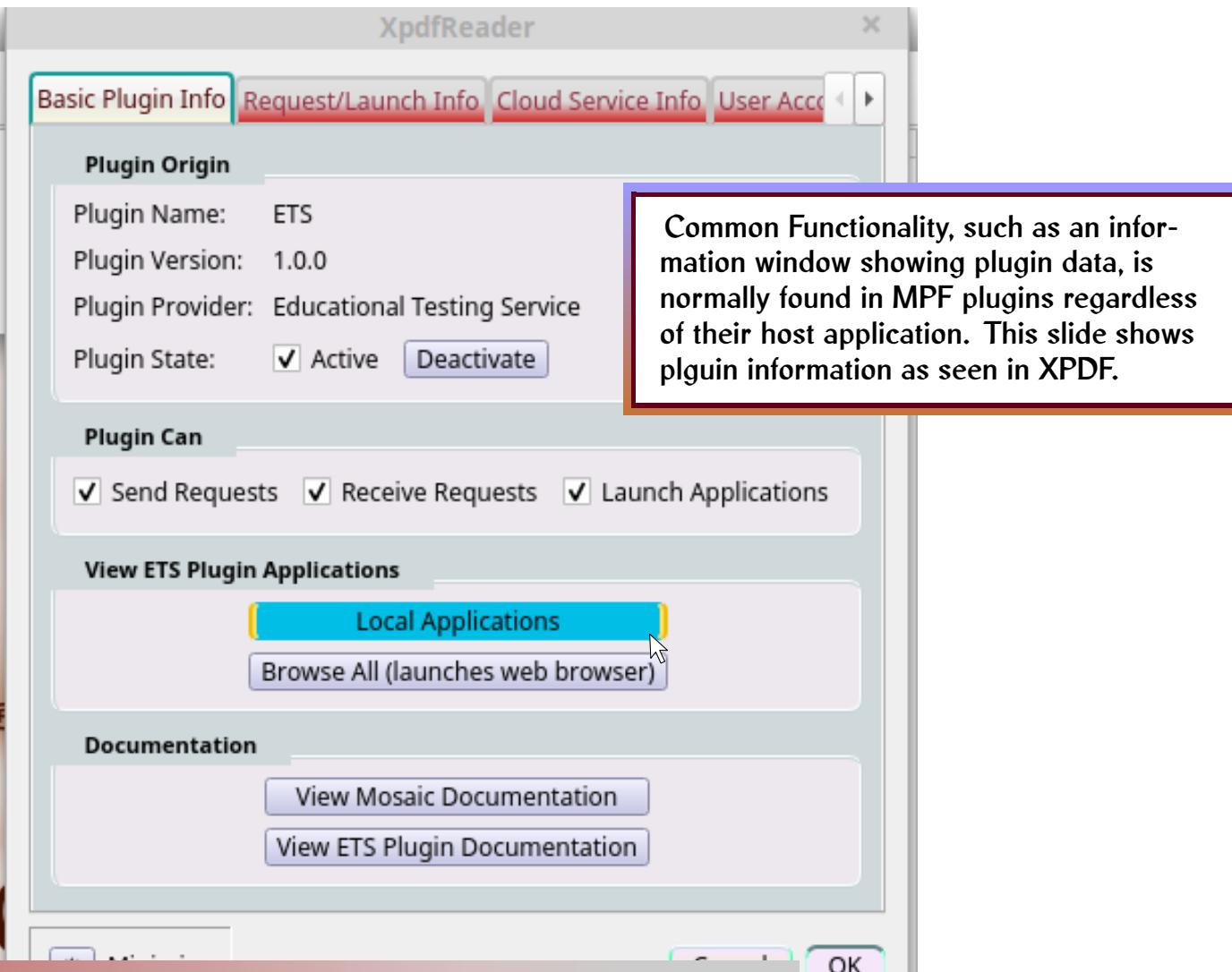
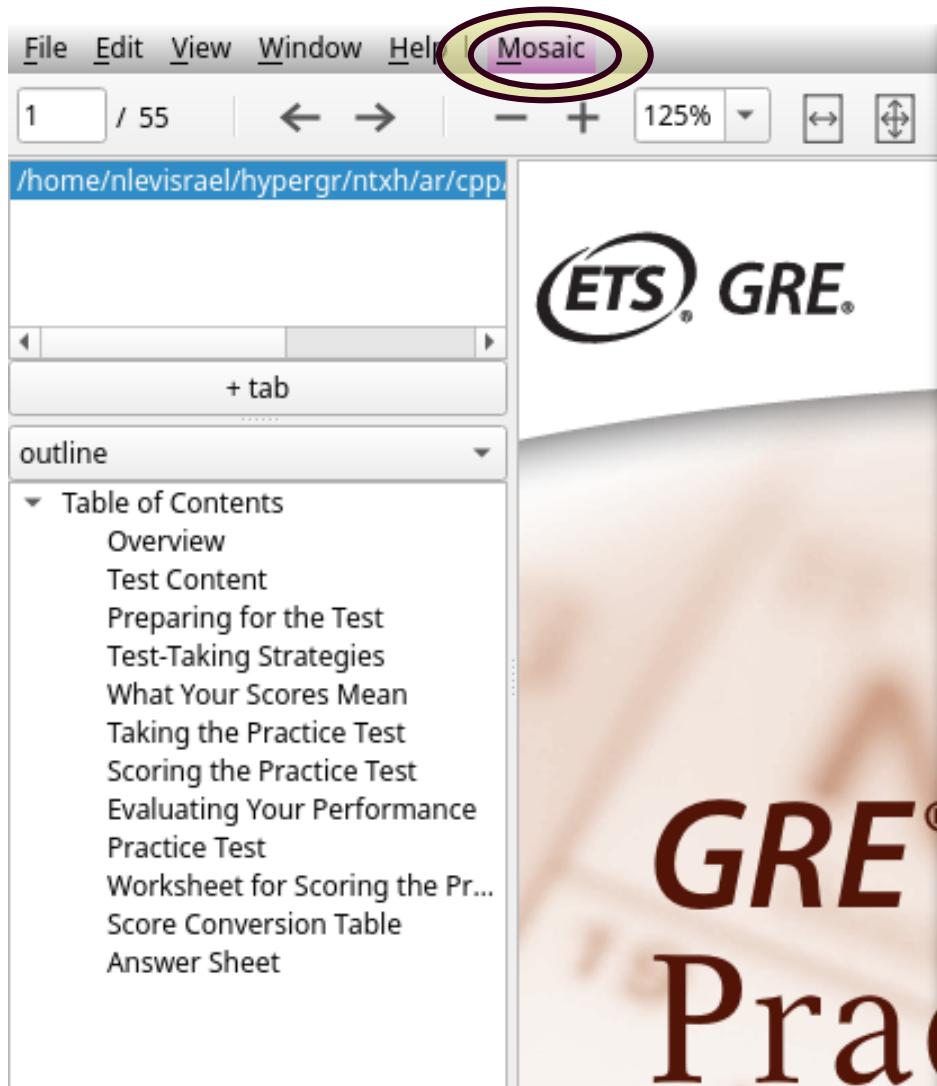
New molecule

IQmol has received data from the student reading a question about thionyl chloride ( $\text{SOCl}_2$ ) and has loaded a Molecular Data file for  $\text{SOCl}_2$ . 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

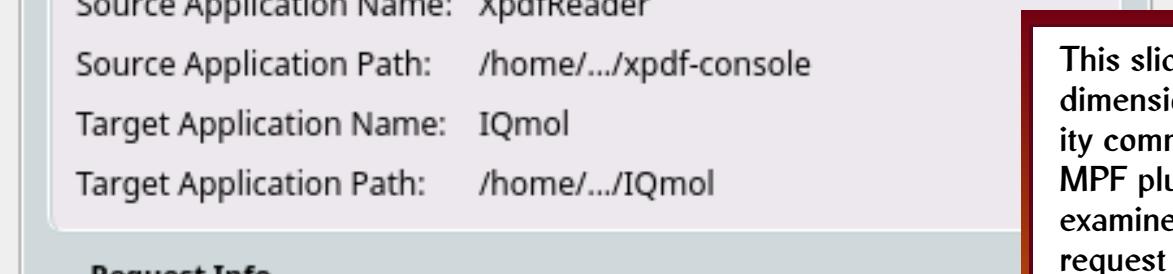
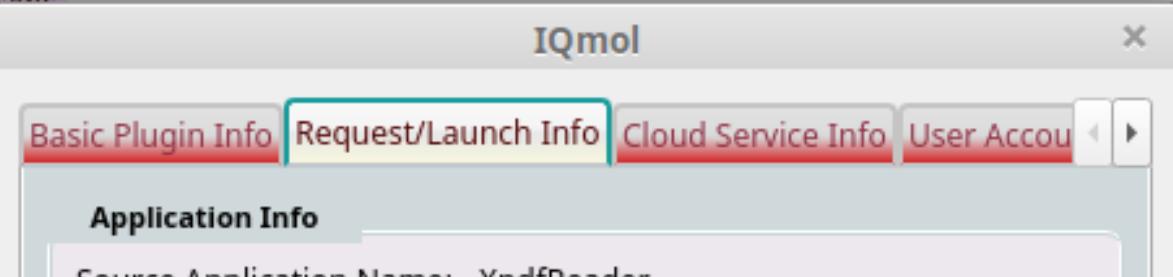
MPF Plugins have similar functionality and features in different host applications, which makes it convenient to use as readers switch among multiple applications.



# MPF Request Info

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

File Edit Display Build Calculation Help | Mosaic



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% | | I | O | find | |

/home/nleviisrael/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  $\text{H}_2\text{O}/\text{H}_2\text{SO}_4$ .
- (E) Lactose has a  $1,1'\text{-}\alpha$ -glycosidic linkage.

97. A peptide digest yields the following fragments listed above. The three fragments are separated using capillary electrophoresis, at which each peptide has a different net charge. Which of the following lists the fragments in the order, from first to last, in which they would reach the detector? (A = glycine; K = lysine)

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

98. In fluorescence spectroscopy,  $(\Phi_f)$  is best defined as the

$\text{H}_3\text{N}^+ \text{---} \text{C}(=\text{O}) \text{---} \text{NH} \text{---} \text{C}(=\text{O}) \text{---} \text{NH} \text{---} \text{C}(=\text{O}) \text{---} \text{CH}_2\text{CO}^-$

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.

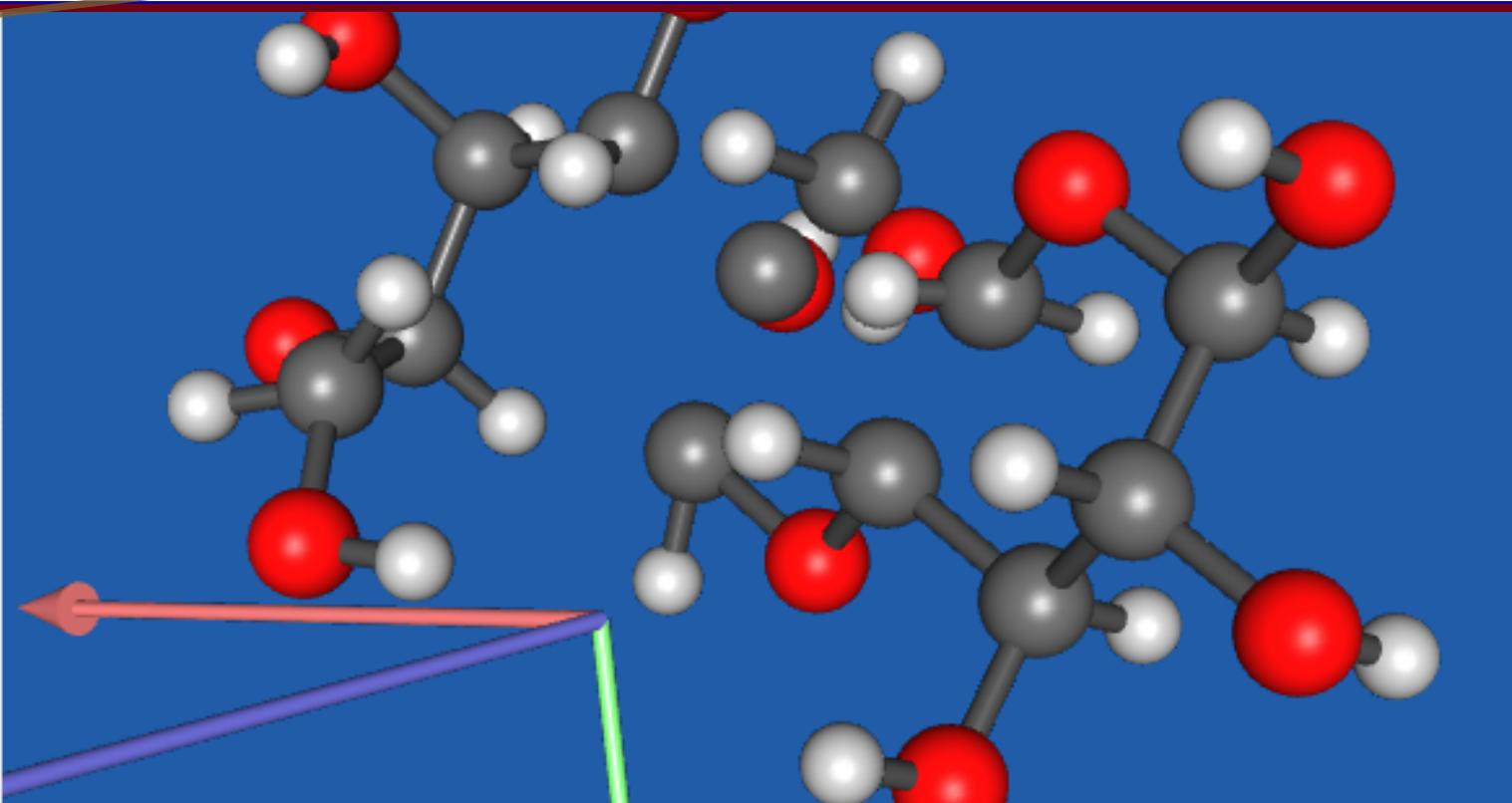
File Edit Display Build Calculation Help | Mosaic



## Model View

- ▶ Global
- ▶  7719-09-7
- ▶  14641-93-1

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).



# Features of Dataset Applications for Books

## Datasets Compiled From Book Examples

The following screenshots demonstrate how data sets can be used even outside of the context of a lab generating experimental data. The pictured data set represents a corpus of linguistic examples mined from the *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 examples can be automated.

**Annotation** Linguistic corpora 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).

The screenshot displays a user interface for creating a dataset from a book. The interface includes the following components:

- Filter Forms** and **Filter Issues** sections with checkboxes for selecting data types (e.g., Text, Dialog, Intonation, Paragraph, Ambiguity, Context, Logic, Scope, Polarity, Belief, Convention, Idioms).
- A text editor showing the sentence: "I have received the e-mail. ?Nevertheless it's in Dutch." with a yellow highlight.
- A list of text samples on the left, with the second item expanded to show its content: "I have received the e-mail. ?Nevertheless it's in Dutch." and its sub-points: "I have received the e-mail. ?Nevertheless it's in Dutch." and "Her husband is in hospital. Yet she's seeing other men.".
- A table on the right showing a list of samples with columns for Form, Text, and Page Number (e.g., Text 22 (N\_A) 257, Text 23 (N\_A) 257, Text 24 (N\_A) 257, Text 25 (N\_A) 257, Text 26 (N\_A) 257, Text 27 (N\_A) 260, Dialog 28 (N\_A) 260).
- Navigation controls at the bottom: **Filtered Up/Down**, **Examples Up/Down**, **Peer Up/Down**, **Chapter Start/End**, **Chapter Up/Down**, **First** (with a circular arrow icon), **Auto Expand** (set to **ON**), **OK**, and **Process**.
- Context menus with numbered callouts (1, 2, 3) highlighting specific features: (1) points to a context menu for a table cell; (2) points to a group of scroll buttons; (3) points to a context menu for a list item.

# Interacting with Data Samples

Filter Forms      Filter Issues

Text       Logic       Scope  
 Inton       Convention       Idioms

**Activate TCP**   **Screenshot**  
**Customize Build**

The linguistic samples comprising this data set are all example sentences, phrases, or dialog-snippets that are used, in the *Blackwell Handbook of Pragmatics*, as expository samples for case-studies of various linguistic phenomenon and pragmatics, semantics, and grammatical theories.

in Dutch.      Show Original      OFF

Text	Form	#	Issue	Page	Chapter
► She was never really happy here. So she's leaving.	Text	19	(N_A)	256	10
► She'll be better off in a new place.	Dialog	20	(N_A)	256	10
► I have received the e-mail,	Text	21	(N_A)	257	10
► I have received the e-mail.	Text	22	(N_A)	257	10
I have received the e-mail.		22		257	10
► Her husband is in hospital.	Text	23	(N_A)	257	10
► Her husband is in hospital	Text	24	(N_A)	257	10
► Her husband is in hospital.	Text	25	(N_A)	257	10
► Her husband is in hospital.	Text	26	(N_A)	257	10
► Oscar knocked the vase ar	Text	27	(N_A)	260	10
► Did Oscar break the vase?	Dialog	28	(N_A)	260	10

Show in Document (requires XPDF)  
Copy Text to Clipboard  
Launch Triple-Link Dialog with Text  
**Copy Samples to Clipboard**  
Highlight (scroll from here)

Filtered Up/Down      Examples Up/Down      Peer Up/Down      Chapter Start/End      Chapter Up/Down      First   
                                
Auto Expand   
OK      Proceed      Close  
Minimize

# Linking Back to the Book

Filter Forms      Filter Issues

Text    Dialog  
 Intonation    Paragraph  
 Ambiguity    Context  
 Polarity    Belief

In France, Watergate wouldn't have done Nixon any harm

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 thought
- ▶ 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 harm
- ▶ In France, Watergate wouldn't have done Nixon any harm
- ▶ The crook paid them with fake money.
- ▶ The crook thought he was paying them with fake money, but
- ▶ We do not know much about this part of the brain, which is

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

Minimize

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

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

File Edit View Window Help  
690 / 867 | ← → | - + | 113% | find |

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

The diagram illustrates a conceptual integration network. At the top, a circle labeled 'Generic space' contains a triangle with three dots. Below it, two circles labeled 'Input I<sub>1</sub>' and 'Input I<sub>2</sub>' also contain triangles with three dots. Dashed lines connect the generic space to each input space. Solid lines connect the two input spaces to a bottom circle labeled 'Blend'. The 'Blend' circle contains a square with three dots, representing the emergent structure.

Figure 29.4 Diagram showing conceptual blending

# A Linguistics Annotation System

## Tools to Facilitate Annotating Linguistic Corpora

The following three screenshots show an example of how a custom-designed application can facilitate the task of building an annotated corpus from a linguistics text. The components demonstrated here enable several strategies (which can be combined) for describing 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.

**Dependency Grammar** Representing phrase structures via inter-word syntactic relationships.

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

# Building Parsing Models

dsmain-console <2>

Filter Issues

Dialog  Ambig.  Polarit.

Paragraph

She has invited at least Sarah and James.

Add (Pair/Triple) Reset

Clear <- ( ( ( -> ) ) -> Copy Read Splice Back Splice

SXPR Mode ( has invited ) 1

She has invited at least Sarah and James.

now much about this part of the  
order shrimp, let alone Louise,  
stor!?  
, but superb.  
order squid?  
have an odd number of books  
et an odd-numbered ticket, let  
get a poor man to wash your  
ed at least Sarah and James.  
nvited at least Sarah and Jam  
no students passed

vn Examples Up/Down

Minimize

OK Proceed Cancel

Link Grammar (Completion Layer)

AAA	AF	AJ	AL	AM	AN	AZ	B	BI	BT
BW	C	CC	CO	CP	CQ	CV	CX	D	DD
DG	DP	DT	E	EA	EB	EC	EE	EF	EI
EL	EN	EP	EQ	ER	EW	EZ	FL	FM	G
GN	H	HA	I	ID	IN	IV	J	JG	J
Q	JT	K	L	LE	LI	M	MF	MG	MJ
MV	MX	N	NA	ND	NF	NI	NJ	NM	NN
NO	NR	NS	NT	NW	O	OD	OF	ON	OT
OX	P	PF	PH	PP	Q	QI	QJ	QU	R
RJ	RS	RW	S	SF	SFI	SI	SJ	SX	SXI
TA	TD	TH	TI	TM	TO	TQ	TR	TS	TT
TW	TY	TZ	U	UN	V	VC	VJ	W	WN
WR	WV	X	XI	Y	YP	YS	Z	ZZZ	

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

The list of link/dependency relations is also isolated as a "dock widget" that may be dragged to float above the other application windows (1), or "docked" at different positions (left or right) on its parent window. This screenshot also shows a dialog box used for a precis of the individual CoNLL-U (Conference on Natural Language Learning - Universal) and Link Grammar relations (2).

1

2

dsmain-console <2>

Dependency Grammar (Refinement)

acl	advcl	advmod
appos	aux	case
ccomp	clf	compound
cop	csubj	dep
discourse	dislocated	expl
flat	goeswith	iobj
mark	nmod	nsubj
obj	obl	orphan
punct	reparandum	root
xcomp		

i nsubj: nominal subject

Ok Hide Details...

A nominal subject (nsubj) is a nominal which is the syntactic subject and the proto-agent of a clause. That is, it is in the position that passes typical grammatical tests for subjecthood, and

Minimize

Issue Page  
(N/A) 698  
(N/A) 699  
(N/A) 700  
(N/A) 700  
(N/A) 702  
(N/A) 703  
(N/A) 703  
(N/A) 704  
(N/A) 704  
(N/A) 704  
(N/A) 704  
Transform  
Signature (macro)

OK Proceed Cancel Proceed

# Link and Dependency Grammar Annotations

dsmain-console <2>

Filter Forms

Text Intonation

She has invited at least Sarah and James

Add at least Reset

SXPR Mode

Clear <- ( ( ( -> ) ) -> Copy Read Splice Back Splice

Text

- We do not know
- Fred won't order
- Him be a doctor
- It's not good, b
- Did Louise order
- She doesn't ha
- She didn't get
- You couldn't ge
- She has invited**  
She has invited at least five

Filtered Up/Down

Minimize

Dependency Grammar (Refinement Layer)

	Pivot	lg:Source Expectation	lg:Target Expectation	lg: Description	dg:Source Expectation
0 {0}	has invited				
1 {1}	<b>invited She</b>				
2 {2}	Sarah James				
3 {3}	at least				

Users can select word-pairs from samples being annotated and then identify the relationship between the selected words, as understood according to Link or Dependency Grammars. The list of link/dependency relations provides an interface to research and read overviews about the relationships.

acl advcl advmod amod  
appos aux case cc  
ccomp clf compound conj  
cop csubj dep det  
discourse dislocated expl fixed  
flat goeswith iobj list  
mark nmod nsubj nummod  
**obj** han parataxis  
punct Unmark vocative  
xcomp Auto Insert

OK Proceed Cancel

Proceed

# Thank You!

Please contact Linguistic Technology Systems for more  
information about Dataset Creator:  
(917) 817-2184  
[amy.neustein@verizon.net](mailto:amy.neustein@verizon.net)

