ADVANCED VISUALIZATION OF SCIENTIFIC METADATA

Ion Mateescu,
Center for International Earth Science Information Network, Columbia University
Lucy Nowell and Leigh Williams, Pacific Northwest National Laboratory*
Karen L. Moe, National Aeronautics and Space Administration

Abstract:

As more and more metadata catalogs become available on line, researchers face the challenge of information overload. Many search tools such as specialized thesauri and more sophisticated query interfaces help users narrow their search when users know specifically what they want. However, in many cases, users are not familiar with what data are potentially available and how different types of data relate to each other. They may need assistance in exploring the contents of one or more distributed data catalogs to better understand the universe of potentially relevant data sets and interrelationships among them. To help alleviate this problem and to increase efficiency in dealing with large metadata collections, the Center for International Earth Science Information Network (CIESIN) at Columbia University is applying research on text visualization to the world of scientific data catalogs. Researchers at Pacific Northwest National Laboratory and CIESIN have worked together to apply the search, visualization and analysis capabilities of WebTheme to collections of metadata documents retrieved using the Z39.50 protocol.

Metadata Catalogs:

Interoperability between different metadata catalogs is facilitated by subscribing to the ANSI/NISO standard Z39.50 which describes a communication protocol for information retrieval. The protocol defines a stateful connection between two machines, in which one machine retrieves records stored in a database on the other machine. In this context the records are metadata records, but they can be bibliographic references or other types of records. The protocol contains facilities to manage the connection, perform searches and store and manipulate the results of the searches. The Z39.50 standard developed in the 1980's has many applications in the library world, and is now at the center of exciting research related to distributed databases, catalog interconnectivity and data mining. Some of the noteworthy new developments in the Z39.50 sector that we are following include:

- Use of W3C standards, such as Resource Description Framework(RDF) and Extensible Markup Language (XML).
- Integration of the functionality of a Z39.50 client into the Mozilla open-source browser.
- Incorporation of SQL as an alternative query language.

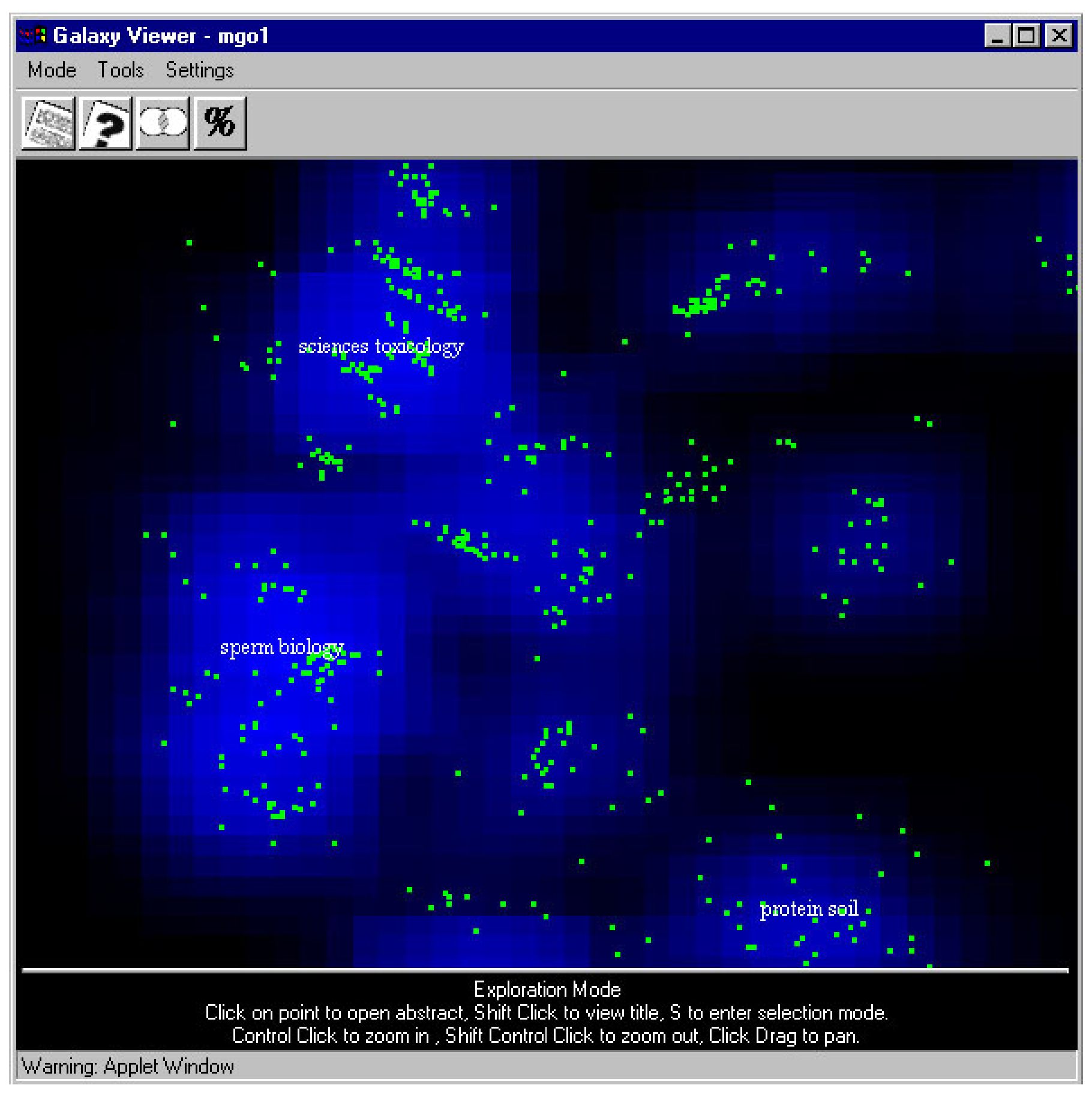
CIESIN has created a custom client program that uses a query language of Reverse Polish Notation (RPN) Type that is understood by computers running Z39.50 server software. A new implementation of the program can be found at http://sedac.ciesin.org/cgi-bin/charlotte?protocol=gw. The client can access any Z39.50 server, but the web implementation is focused on providing access to a number of known metadata catalogs for the Earth Science community.

Introduction:

In the past ten years, the Internet explosion has caused a millionfold increase in the diameter of our information universe. In this universe, if the content similarity of different documents were visualized as a form of gravity, it would provide a limited structure and orientation. While documents are amorphous objects, whose structure is intrinsic to their family of documents as far as this universe is concerned, metadata is a structured way of looking at pieces of information, textual or otherwise. Metadata is addressing the lack of structure in the data - it is structured data about data. It improves search precision and by utilizing standard thesauri, it minimizes language effects due to polysemy, synonymy and ambiguity in search and classification. WebTheme is a software application developed by Pacific Northwest National Laboratory (PNNL) that measures interdocument similarity, which it represents as proximity in a two dimensional Galaxies visualization.

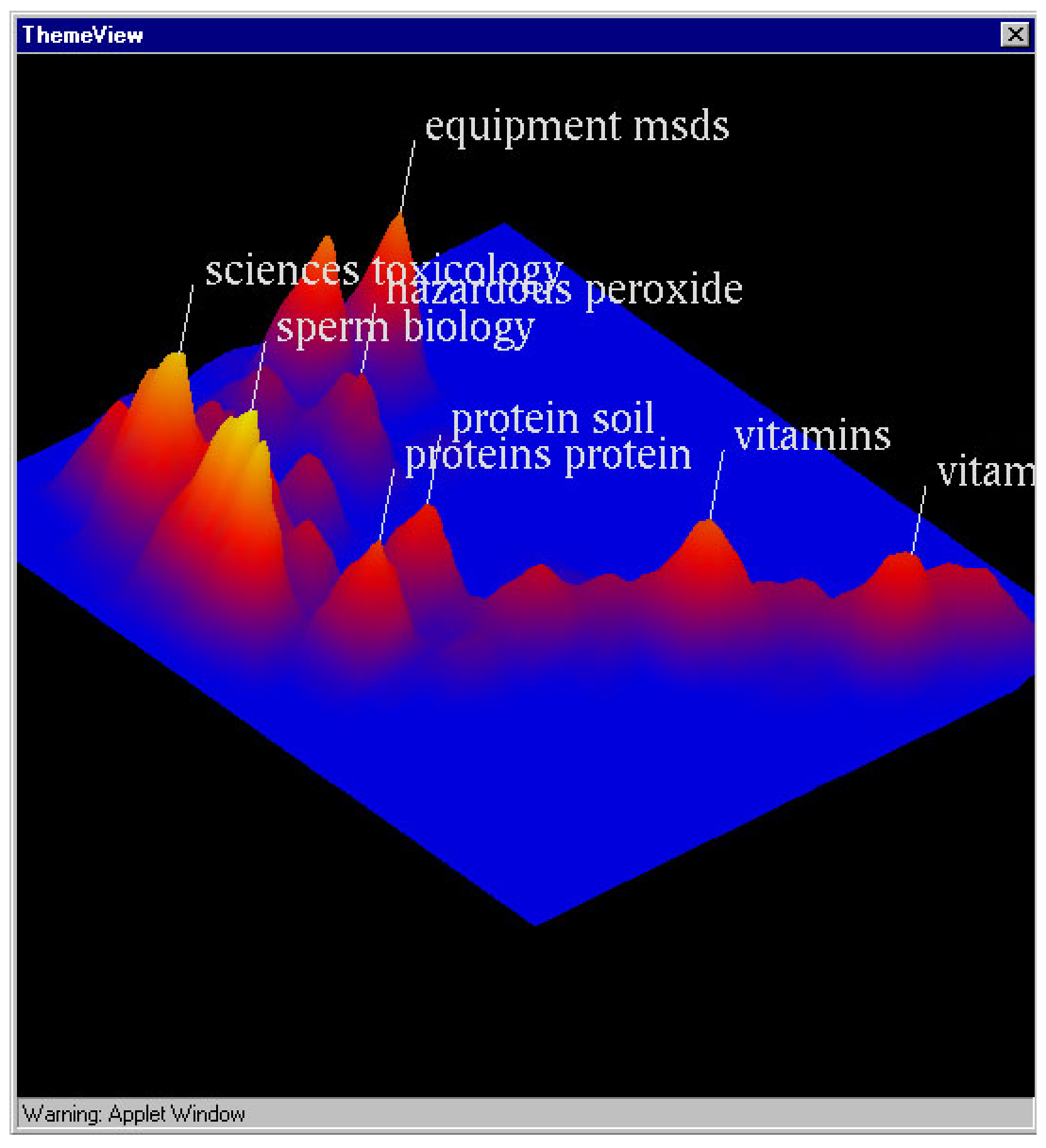
WebTheme:

Producing a WebTheme text visualization is a three stage process: harvest, text analysis and visualization. Given a list of keywords, the WebTheme uses commercial search engines to retrieve related documents from the web. It can also harvest from a list of URLs provided by the user. The harvested HTML files are converted to text and analyzed and clustered based on their content. Two clustering algorithms can be used: K-mean and Hierarchical Clustering. The results are stored on the server and can be accessed anytime with a browser. WebTheme provides Java applets for visualizing the analyzed collections that it retrieved from the Web: a constellation-like Galaxies visualization in which distance is a measure of similarity and a ThemeView visualization in which peak height reflects topic strength. Both visualizations allow zooming, panning and inquiries.



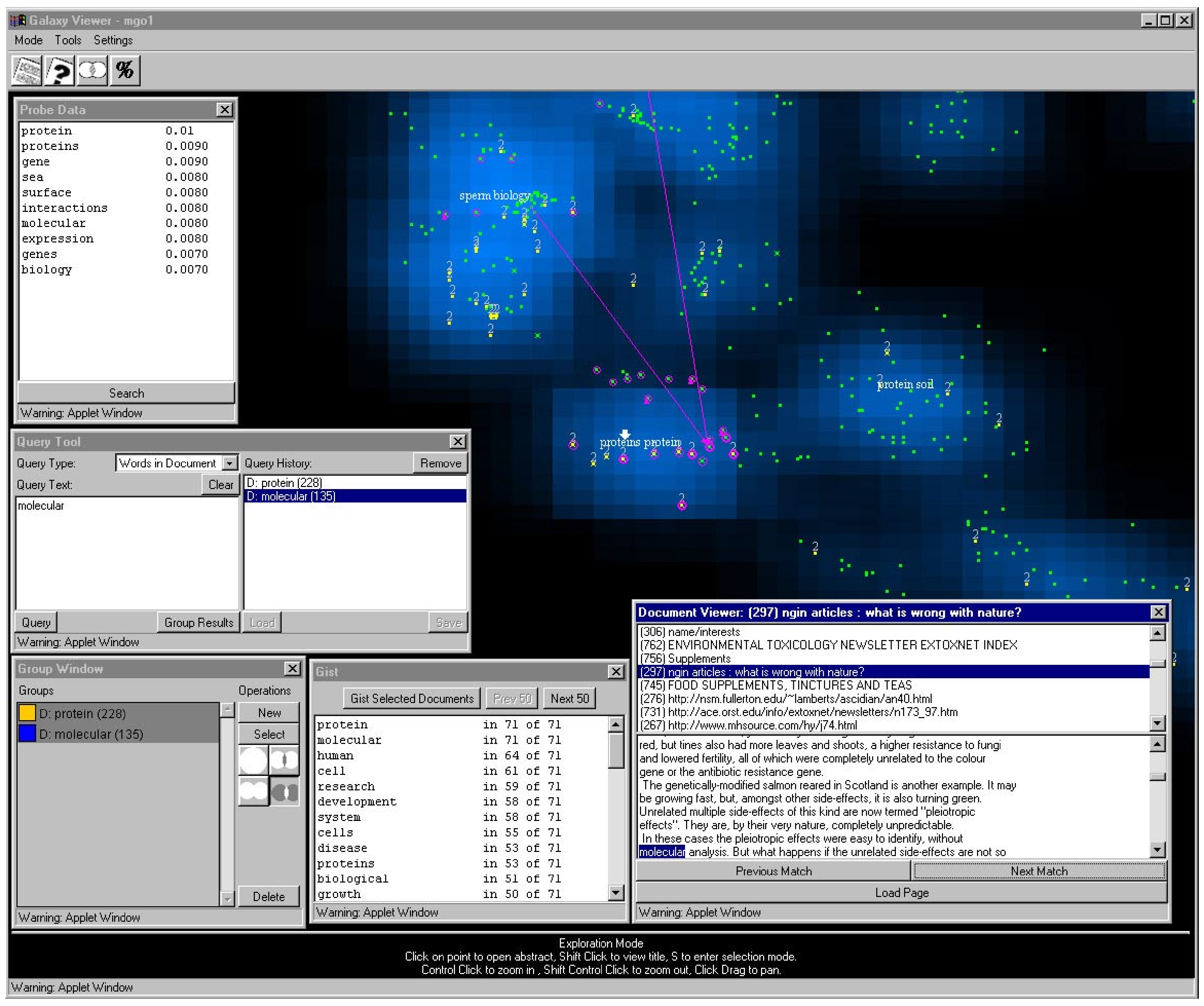
Exploration Mode
Click on point to open abstract, Shift Click to view title, S to enter selection mode.
Control Click to zoom in , Shift Control Click to zoom out, Click Drag to pan.

Warning: Applet Window



Warning: Applet Window

Webtheme provides a suite of analysis tools written in Java that allow the user to explore documents and the links between them. The query tool allows the user to find the documents containing search terms, the Group tool provides set operations and color coding for selected documents. The Gist tool shows a ranked list of main topics in selected documents. The Galaxies View and ThemeView visualizations provide the user a visual appraisal of the distance in topicality between the texts as well as a view of the document location in the general knowledge map. In Link mode one can see hypertext coreferences between documents. The Document Viewer loads the content of the selected documents and facilitates navigating the collection of documents



Probe Data

```
protein        0.01
proteins       0.0090
gene           0.0090
sea            0.0080
surface        0.0080
interactions   0.0080
molecular      0.0080
expression     0.0080
genes          0.0070
biology        0.0070
```

Warning: Applet Window

Query Tool
Query Type: Words in Document
Query Text: molecular
Query History:
D: protein (228)
D: molecular (135)

Warning: Applet Window

Group Window
Groups
Operations
🟨 D: protein (228)
🟦 D: molecular (135)

Warning: Applet Window

Gist

```
protein        in 71 of 71
molecular      in 71 of 71
human          in 64 of 71
cell           in 61 of 71
research       in 59 of 71
development    in 58 of 71
system         in 58 of 71
cells          in 55 of 71
disease        in 53 of 71
proteins       in 53 of 71
biological     in 51 of 71
growth         in 50 of 71
```

Warning: Applet Window

Document Viewer: (297) ngin articles : what is wrong with nature?

(306) name/interests
(762) ENVIRONMENTAL TOXICOLOGY NEWSLETTER EXTOXNET INDEX
(756) Supplements
(297) ngin articles : what is wrong with nature?
(745) FOOD SUPPLEMENTS, TINCTURES AND TEAS
(276) http://nsm.fullerton.edu/~lamberts/ascidian/an40.html
(731) http://ace.orst.edu/info/extoxnet/newsletters/n173_97.htm
(267) http://www.mhsource.com/hy/j74.html

red, but tines also had more leaves and shoots, a higher resistance to fungi and lowered fertility, all of which were completely unrelated to the colour gene or the antibiotic resistance gene.
The genetically-modified salmon reared in Scotland is another example. It may be growing fast, but, amongst other side-effects, it is also turning green. Unrelated multiple side-effects of this kind are now termed "pleiotropic effects". They are, by their very nature, completely unpredictable.
In these cases the pleiotropic effects were easy to identify, without molecular analysis. But what happens if the unrelated side-effects are not so

Warning: Applet Window

Exploration Mode
Click on point to open abstract, Shift Click to view title, S to enter selection mode.
Control Click to zoom in , Shift Control Click to zoom out, Click Drag to pan.

Warning: Applet Window

Evaluation:

Although the WebTheme technology looks very promising, several problem areas have been identified:

- The structured format of a metadata record makes it more likely that local formatting standards will have an undue influence on the clustering results. Our experience shows that at this stage the WebTheme software performs better with unstructured documents.
- The abstract in a metadata record is the most relevant part for analysis purposes because it refers directly to the content of the data. However, abstracts are short by definition - under 200 words - and the discriminant function does not perform as well as we would like.
- The ability to refine the results of the web search is currently limited.
- There are major differences between the HTTP protocol and the Z39.50 protocol related to both the connection setup and the format of the data that is retrieved.

Some of these issue are being addressed by developers at PNNL and CIESIN. New developments in text mining technologies and Z39.50 standards and applications are being monitored and may find their way into future implementations.

Acknowledgements
The views expressed in this paper are those of the authors and do not necessarily represent the views of CIESIN, Columbia University, PNNL, Battelle or the sponsors. Support for this paper was provided by U.S. National Aeronautics and Space Administration under Contract NAS5-98162.

* Pacific Nortwest National Laboratory is managed by Battelle Memorial Institute for the United States Department of Energy under Contract DE-AC06-76RLO 1830.