MetaViewer and MetaCommander: Applying WWW Tools to Genome Informatics

Yasuhiko Kitamura [1]
kitamura@info.eng.osaka-cu.ac.jp
Hideyuki Nakanishi [2]
nuka@lab7.kuis.kyoto-u.ac.jp
Tetsuya Nozaki [1]
noz@kdel.info.eng.osaka-cu.ac.jp
Teruhisa Miura [2]
miura@lab7.kuis.kyoto-u.ac.jp
Toru Ishida [2]
ishida@kuis.kyoto-u.ac.jp

[1] Department of Information and Communication Engineering,
Faculty of Engineering, Osaka City University
3-3-138 Sugimoto, Sumiyoshi-ku, Osaka 558, Japan

[2] Department of Information Science,
Faculty of Engineering, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-01, Japan

Abstract

With the advance of the Human Genome Project, a huge amount of various genome data has been stored in a number of databases and the WWW system is widely used to access these databases. From the viewpoint of information supplier, the WWW is a quite useful tool to provide various types of data easily, but from the viewpoint of information consumer, it is not good enough because of lack of rigid data format and difficulty of data access. In this paper, by extending a current WWW browser, we propose two generic WWW tools; MetaViewer and MetaCommander, and try to apply them to the genome informatics to support researchers who search, analyze, and dispatch genome data, and discuss their potential advantages from the viewpoint of information consumer.

1 Introduction

With the advance of the Human Genome Project, a huge amount of various genome data has been stored in a number of databases and the amount is rapidly increasing day by day. Most contents of current genome databases are represented as text data such as nucleic acid or amino acid sequences with their bibliographical data, and furthermore graphical data are also going to be used widely to represent more complex structures such as motif, pathway, or 3D-structure of protein. Nowadays, with advance and pervasion of information network technology, major genome databases are linked together by using the WWW interface to be offered to researchers all over the world.

The WWW technology has been rapidly pervaded on the Internet because of its high freedom of data format and its easiness of information dispatching. Information suppliers can make their own homepages as easy as they do word-processing and their homepages can contain not only text but also graphic, sound, and movie data. Furthermore, it is easy to relate a page with other pages on the Internet by attaching hyperlinks. An information supplier can dispatch his/her own information toward the world just by creating his/her homepages on a server machine connected to the Internet.

From the viewpoint of information consumer, however, it is not easy to find desired pages or to analyze contained data mechanically since the contained data are not sorted out in a standard manner because of its high freedom of data format. Complex hyperlinks make finding desired pages more difficult and often makes novice users get lost in the web. The current WWW system hence is a useful tool for information suppliers but not good enough for information consumers who collect data, reorganize them, and dispatch them again.

The Human Genome Project tends to gradually move its research effort from structure analysis such as nucleic acid or amino acid sequences alignment to function analysis such that estimates functions of enzymes or proteins. Function analysis is a challenging research field, but its methodology has not been established, so we often have to take trial and error approaches. Hence it is essential to make effective use of facts already found and compiled in genome databases to verify hypotheses of function or to reduce the number of hypotheses. A system will be appreciated which supports researchers to collect data from a number of various databases, compare them, classify them, and dispatch the obtained result immediately to other researchers.

In this paper, by extending the current WWW technology, we propose two generic WWW tools; MetaViewer and MetaCommander, and try to apply them to the genome informatics to support researchers who search, analyze, and dispatch genome data, and discuss their potential advantages from the viewpoint of information consumer.

2 MetaViewer

2.1 Background

How we retrieve only desired data and present them effectively is one of fundamental research issues of large-scale database systems. The WWW system stores a huge amount of data, but has no rigid data model and data management system, so it leaves the above issue to its users. However, existing WWW browsers are not designed well enough to support such users. For example, it is not easy to even view or compare data from multiple pages at a glance. Some browsers may open multiple windows, but each window displays the whole page although the user hopes to view just a part of it or the relation among the parts. As a remedy, we propose the MetaViewer which is an interactive WWW tool to collect, arrange, view, store, and dispatch

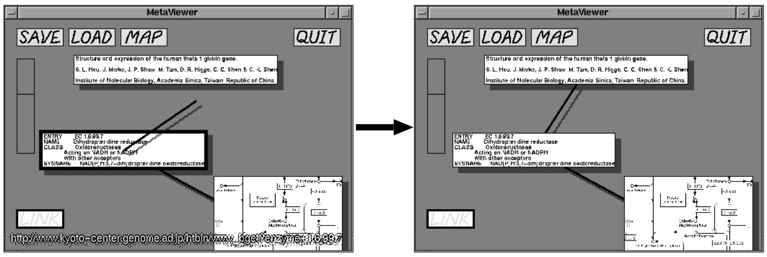
WWW data originated from multiple pages, by extending a current WWW browser. Ohmi et al. [4] also take a similar approach and develop a WWW tool called PAN-WWW, which uses index-card metaphor.

How we sort or arrange retrieved data is also another important research issue to support users' problem solving. We do not discuss much about this issue in this paper, but we can find some papers in the research field called "Creative Thinking Support Systems" [3].

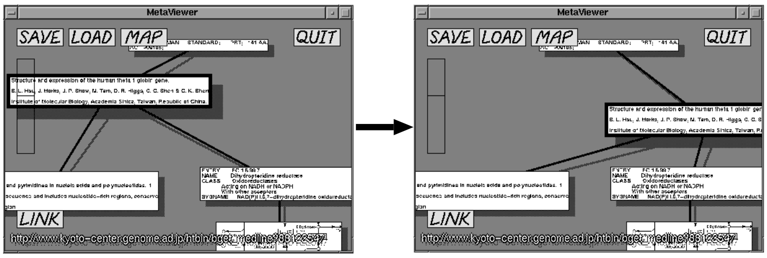
2.2 Functions

Functions of MetaViewer are described as follows.

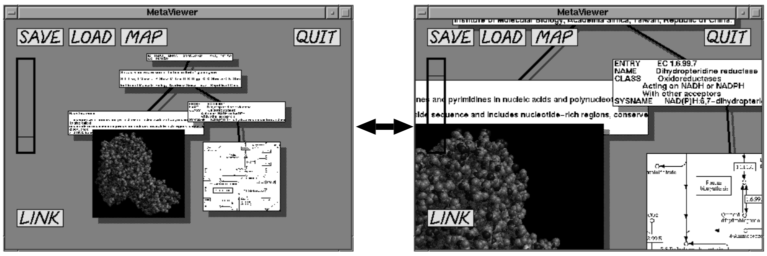
Collecting data. We can easily collect pieces of text or graphical data through a WWW browser by cutting and pasting it onto the MetaViewer. Collected pieces of data are represented as a group of panels on the field of MetaViewer If we like to emphasize a relation between two panels, we can attach a link between them as shown in Figure 1(A).



(A) Making a link

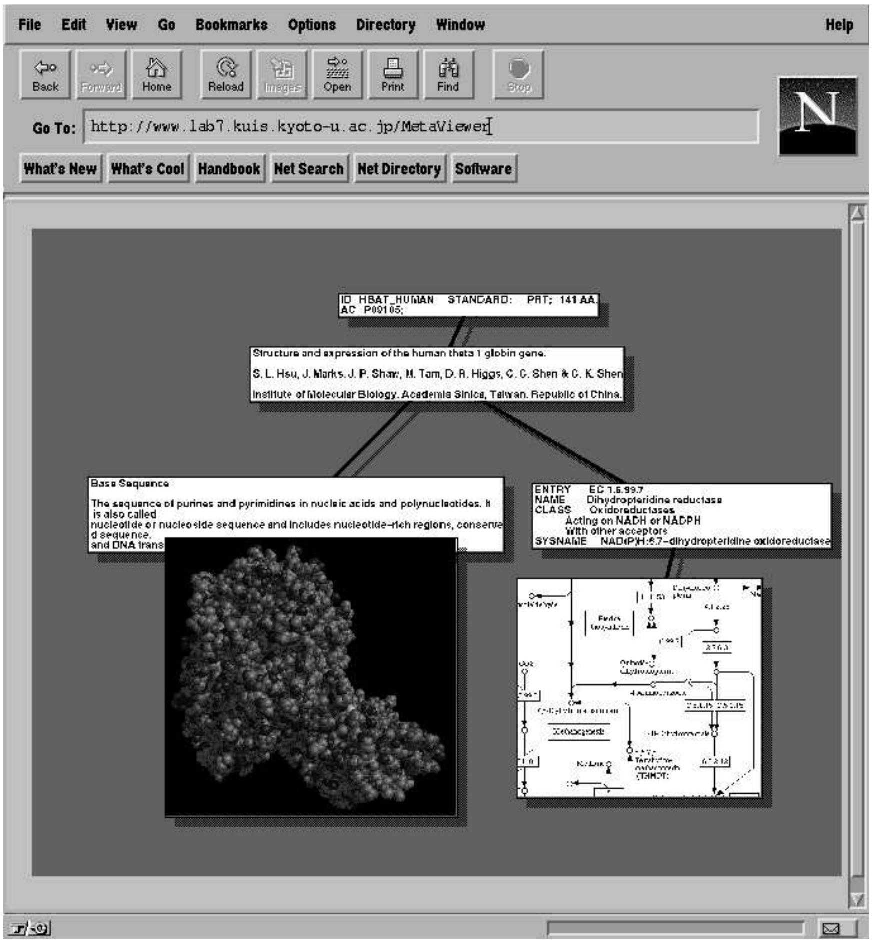


(B) Moving a panel



(C) Zooming in/out

Figure 1: Operations on MetaViewer.



Go To: http://www.lab7.kuis.kyoto-u.ac.jp/MetaViewer

ID HBAT_HUMAN STANDARD: PRT; 141 AA.
AC P09105;

Structure and expression of the human theta 1 globin gene.

S. L. Hsu, J. Marks, J. P. Shaw, M. Tam, D. R. Higgs, C. C. Shen & C. K. Shen

Institute of Molecular Biology, Academia Sinica, Taiwan, Republic of China.

Base Sequence

The sequence of purines and pyrimidines in nucleic acids and polynucleotides. It
is also called
nucleotide or nucleoside sequence and includes nucleotide-rich regions, conserve
d sequence.
and DNA trans

ENTRY EC 1.6.99.7
NAME Dihydropteridine reductase
CLASS Oxidoreductases
Acting on NADH or NADPH
With other acceptors
SYSNAME NAD(P)H:6,7-dihydropteridine oxidoreductase

Figure 2: An example of clickable map.

Rearranging panels. We can rearrange panels as we like shown in Figure 1(B). Note the attached links move with the panel.

Zooming in/out. We can zoom in/out the field of the MetaViewer as shown in Figure 1(C). By zooming in the field, we can view details of the panels, and by zooming out, we can view the whole group of panels at a glance.

Linking panels with originated pages. By clicking a panel, we can make the WWW browser show the WWW page from which the data of the panel is originated.

Storing data. We can save the field of the MetaViewer into a file.

Dispatching data. We can transform the field of the MetaViewer into a clickable map in the HTML format as shown in Figure 2, and can dispatch the results obtained through operations on the MetaViewer.

2.3 Example

We can collect text and graphical data concerning the human hemoglobin from various databases as shown in Figure 2. These contents are originated from the following databases.

- Descriptions of the hemoglobin theta-1 chain from a protein database SWISS-PROT (P09105).
- Bibliographical data about the hemoglobin theta-1 chain from MEDLINE (MUID:88122547) and descriptions about "base sequence".
- Descriptions about the dihydropteridine reductase from ENZYME (EC 1.6.99.7).
- Pathway data around the above enzyme from PATHWAY database (MAP00790).
- 3D-structure data of hemoglobin from NRL_3D database at the John Hopkins University (`http://www.gdb.org`).

2.4 Applications to genome informatics

As we mention above, the WWW interface has been widely used to access various genome databases. However, from the viewpoint of information consumer, we can point out several shortcomings in current WWW browsers and the MetaViewer can be a remedy.

- For genome data analysis, we often need to refer a number of WWW pages at once. Current WWW browsers are not designed to be good enough to compare even two pages or to view a number of pages at a glance. By using the MetaViewer, we can cut and paste pieces of desired data from a number of WWW pages and compare or view them on the MetaViewer.
- Hyperlinks to relate WWW pages are usually offered by only information suppliers who made the pages, and it is impossible to add personal links on them. By using the MetaViewer, we can easily make a personal page which contains related WWW pages and links.
- Some research groups are creating databases which represent interactions among molecules such as pathway database[5]. They usually draw a number of clickable maps manually to represent the interactions. We can facilitate such processes by attaching the drawing function to the MetaViewer.

3 MetaCommander

3.1 Background

Another user support in the WWW system is automating information retrieval. There are a number of search engines developed to find desired WWW pages. Most of them use keyword search technique aiming at general purpose use, but recently we are often annoyed with the fact that they often hit a large number of pages with almost the same number of junk pages.

Aiming at special purpose search engines, Etzioni [1] is developing Internet Softbots (Software Robots), for example, the Ahoy! softbot for searching people's home pages and the Shopbot for supporting user's shopping on the Internet.

In this paper, we develop a tool and a language to build special-purpose softbots easily. The MetaCommander is a script-based WWW tool to collect, arrange, store, and dispatch WWW data automatically, by extending a current WWW browser. The MetaCommander consists of an interpreter and a script language. The MetaCommander interpreter interprets a script and executes the commands. User can build their own softbot by describing a script for its purpose.

3.2 Functions

Major commands which are supported by the MetaCommander script language are as follows.

- Open URL:

  `getURL(`*url, filter*`)`
  `postURL(`*url, filter, parameter1, parameter2, ...*`)`

  These commands are used to get an HTML document from the designated URL *url* and to put it into the MetaCommander buffer. The `postURL` command is used for supporting the CGI interface. *filter* is used to transform a non-standard document to a standard HTML format. Each *parameter* is described as "argument_name = strings" and used to post the parameter to the CGI interface.

- Print:

  `print(`*strings*`)`

  This command is for printing out strings *strings*. When *strings* are missing, the document section in the scope is printed out.

- Search strings:

  `searchString(`*number, strings*`)`

  This command is for searching the designated *strings* in the scope. The *number* is set to be the number of repeats.

- Search anchor:

  `searchAnchor(`*number*`)`

  This command is for finding the *number*-th anchor URL in the scope.

- HTML tag:

  `tag(`*tag*`)`

  This command is for inserting a *tag* as we can shape up the output in an HTML format.

```
1:$keyword = "interleukin human"
2:$searchNumber = 5

3:tag( "H1" ) {
4:    print("Keyword(s):" $keyword )}

5:postURL( "http://www.genome.ad.jp/htbin/www_bfind_sub?genbank-today", none,
      "keywords"=$keyword , "selection"="bfind" ) {
6:    tag( "UL" ) {
7:        searchString( $searchNumber , "gb:" ) {
8:            print( "<LI>" )
9:            print()
10:           getURL( searchAnchor( 1 ), filter ) {
11:               searchString( 1, "REFERENCE" ) {
12:               print()
13:               }}}}}
```

Figure 3: An example of the MetaCommander script.

Each command has a scope where it works. For example, let us use the following script.

```
getURL("http://www.kdel.info.eng.osaka-cu.ac.jp/"){
    print()
}
print()
```

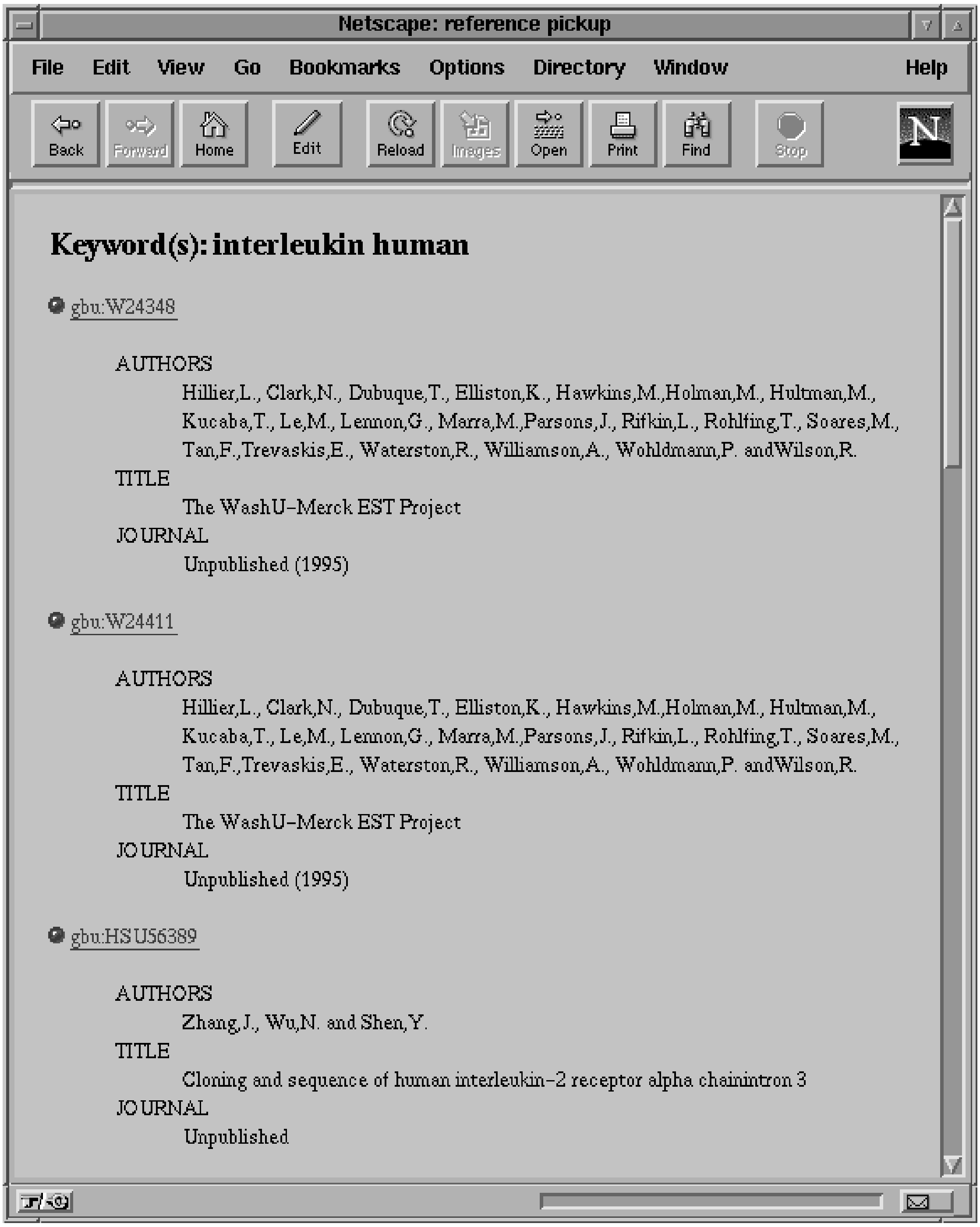
As the first `print` command is in a scope of `getURL("http://www.kdel.info.eng.osaka-cu.ac.jp/")`, it prints out the whole page at the URL. The second `print` command is out of the scope, so it prints out nothing.

3.3 Example

We here give an example of the MetaCommander script in Figure 3. In line 5, the MetaCommander gets a HTML document from GenBank (`http://www.genome.ad.jp/htbin/www_bfind_sub?genbank-today`) with keywords "interleukin human" as defined in line 1. In line 7, it searches strings "gb:" in the buffer and prints out them in line 9. It then searches and opens the first anchor URL (line 10), and prints out the "REFERENCE" section (line 11). We use a filter to transform the obtained page because GenBank pages are not in the standard HTML format. Lines between 8 and 13 are repeated 5 times as defined in line 2 and 7. Lines 3, 6, and 8 are included for the output should be printed as in an HTML format. The output of executing this script is shown in Figure 4.

3.4 Applications to genome informatics

The MetaCommander also can contribute to genome informatics at the following points.



Keyword(s): interleukin human

- gbu:W24348

  AUTHORS
  Hillier,L., Clark,N., Dubuque,T., Elliston,K., Hawkins,M.,Holman,M., Hultman,M., Kucaba,T., Le,M., Lennon,G., Marra,M.,Parsons,J., Rifkin,L., Rohlfing,T., Soares,M., Tan,F.,Trevaskis,E., Waterston,R., Williamson,A., Wohldmann,P. andWilson,R.
  TITLE
  The WashU-Merck EST Project
  JOURNAL
  Unpublished (1995)

- gbu:W24411

  AUTHORS
  Hillier,L., Clark,N., Dubuque,T., Elliston,K., Hawkins,M.,Holman,M., Hultman,M., Kucaba,T., Le,M., Lennon,G., Marra,M.,Parsons,J., Rifkin,L., Rohlfing,T., Soares,M., Tan,F.,Trevaskis,E., Waterston,R., Williamson,A., Wohldmann,P. andWilson,R.
  TITLE
  The WashU-Merck EST Project
  JOURNAL
  Unpublished (1995)

- gbu:HSU56389

  AUTHORS
  Zhang,J., Wu,N. and Shen,Y.
  TITLE
  Cloning and sequence of human interleukin-2 receptor alpha chainintron 3
  JOURNAL
  Unpublished

Figure 4: An output of the MetaCommander script in Figure 3.

Automating data collection. For genome data analysis, we often require to collect a large number of data from various databases. Users which use only interactive WWW browsers have to collect them manually and take long time especially when the network is congested. By using the MetaCommander, this process can be automated. For example, if we prepare a script which describes how to collect data before we leave our laboratory in the evening, the MetaCommander, on behalf of us, collects and sorts the data in the night time, and we can get the sorted data in the next morning.

Automating repeated operations. When we execute the homology search on many molecular sequences, we are often forced to do repeated operations many times. What is worse, if we use a sequence interpretation tool such as BLAST, we are forced to input many parameters at each time. The MetaCommander can automate these operations.

Updating personal databases. The contents of genome databases are updated day by day. Even if we collect data from such databases and make our own personal database for special purpose, it will be obsolete soon. The MetaCommander can access the original databases regularly and can automatically update the personal databases to the latest version.

4 Summary and Future Study

We introduced the MetaViewer and the MetaCommander which are generic WWW tools for supporting information consumers and have shown how they can be applied to genome informatics. These are under development on the UNIX operating system by extending a current WWW browser as shown in Figure 5. We are planning to release beta versions of them available to genome researchers near future. Collecting comments from them, we evaluate and further extend these tools. Our current future study includes the following points.

Automatic generation of script. It is not easy for novice users to make the MetaCommander scripts from scratch by themselves. It is possible to generate the scripts semi-automatically by integrating the MetaViewer and the MetaCommander. Namely, we can extend the MetaViewer as it generates a MetaCommander script which reflects the history of usage. Users can change the script also as they like by using an text editor.

Integrating inference scheme. To estimate functions of genes or proteins, it is required to not only collect a large number of data, but also verify hypotheses or reduce the number of them. By integrating AI techniques such as search, deduction, and learning capabilities into the MetaCommander, it will serve genome informatics as a more advanced analysis tool.

Acknowledgement

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, “Genome Science,” from The Ministry of Education, Science, Sports and Culture in Japan.

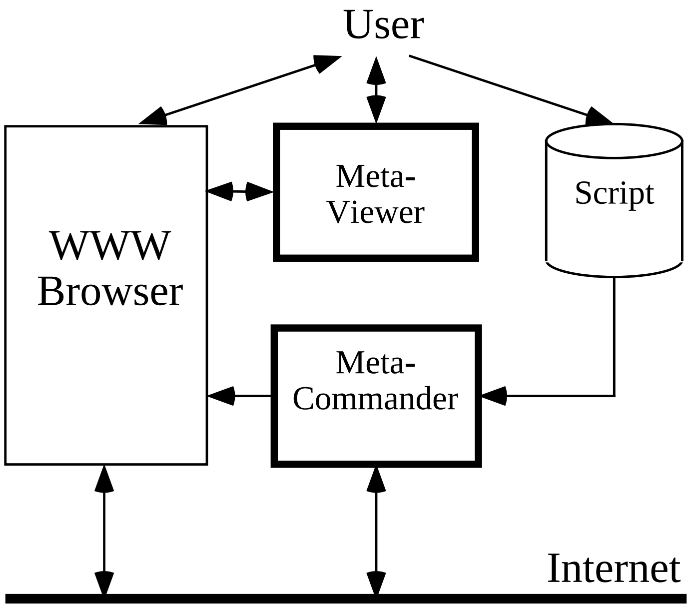


Figure 5: Implementation of MetaViewer and MetaCommander.

References

[1] Oren Etzioni, "Moving Up the Information Food Chain: Deploying Softbots on the World Wide Web," Proceedings Thirteenth National Conference on Artificial Intelligence, pp.1322–1326, 1996.

[2] Minoru Kaneshisa, "Genome joho eno shotai (in Japanese)," Kyoritsu Shuppan, 1996.

[3] Susumu Kunifuji, "A Survey on Creative Thinking Support Systems and the Issues for Developing Them," Journal of Japanese Society for Artificial Intelligence, Vol.8, No.5, pp.552–559, 1993.

[4] Yoshihiro Ohmi, et al., "PAN-WWW: A Card-Based Internet Resource Accessing Tool," Transactions of Information Processing Society of Japan, Vol.37, No.1, pp.154–162, 1996.

[5] Naoko Tateishi, et al., "An Integrated Database SPAD for Signal Transduction and Genetic Information," Proceedings Genome Information Workshop 1995, pp.160–161, 1995.