

University of Illinois  
at Urbana-Champaign

National Center for  
Supercomputing Applications

152 Computing Applications Building  
605 East Springfield Avenue  
Champaign, IL 61820

217 244-0072

Michael D. Doyle, Ph.D.  
Director  
Center for Knowledge Management  
University of California, San Francisco  
530 Parnassus Avenue, box 0840  
San Francisco, CA 94143-0840

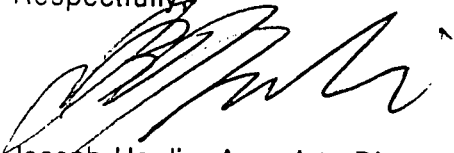
January 21, 1994

Michael:

This letter will serve as notification that SDG is pleased to collaborate with you on the Digital Library project "Embedded Visualization Object for Knowledge Access, Creation and Management through the World Wide Web." That scope of work for NCSA will focus on extensions of NCSA Mosaic and development of URN/URC methods.

We look forward to working with you on this proposed project.

Respectfully,



Joseph Hardin, Associate Director  
Software Development Group

## NSF/ARPA/NASA Digital Libraries RFP Response

**Title: A Knowledge Management Environment through the World Wide Web**

**Principal Investigator: Michael D. Doyle, Ph.D., UCSF Library and Center for Knowledge Management**

### Specific Aims:

1) To develop a prototype knowledge management environment for the biomedical sciences which integrates access to online representations of the scientific literature, bibliographic databases, high-performance visualization technologies, large-scale scientific databases, and tools for authoring new-generation scientific publications.

1.a) To explore and evaluate the applicability of these tools in the areas of radiology and developmental & molecular biology.

2) To provide a means for relating digital forms of spatial, functional, and conceptual information as a basis for linking the biomedical scientific literature, through the Red Sage electronic journals project, to data resources provided through the Visible Human Project, The Human Brain Project, The Visible Embryo Project, The Human Genome Project, The Protein Database, and other large-scale biomolecular and biostructural databases.

2.a) To exploit these linking strategies in the creation of a set of integrated semi-automatic front ends to varied scientific databases accessible through the Internet.

2.b) To incorporate these linking methodologies into interactive authoring and editorial tools, allowing the creation of online publications that can embed visualizations and simulations which draw data from these Internet-accessible scientific databases.

3) To develop tools which provide access to interactive visualization and analysis of massive biomedical datasets through the Internet's World Wide Web distributed hypermedia network.

3.a) To refine and extend our existing algorithms enabling distributed visualization and analysis software "engines" which can be efficiently accessed by remote users through the Internet.

3.b) To refine and extend our existing algorithms to allow the display and real-time interactive control of three- and four-dimensional data visualization and analysis tools within hypermedia documents viewed using NCSA's Mosaic graphical browser to the World Wide Web.

3.c) To develop algorithms which use novel compression technologies for the optimized interactive remote control of computationally-intensive graphical applications through the Internet.

3.d) To integrate a, b & c into a system which allows real-time remote access to distributed parallel computational applications for visualization and analysis resources within a distributed hypermedia environment.

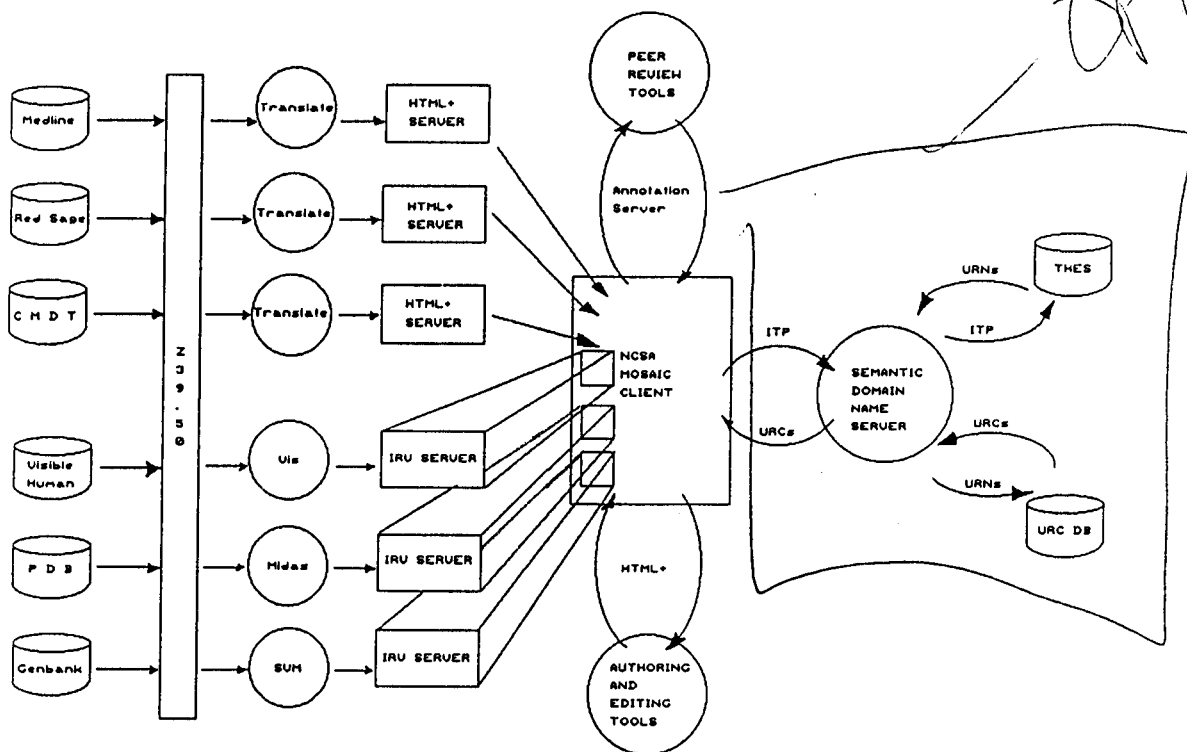
4) To explore extensions of the paradigm of scientific publishing which are made possible through use of current multimedia technologies in a networked environment, including:

4.a) publishing multidimensional datasets integrated with articles, eg: MRI and molecular data, preferred views, animations, interactive visualizations, interactive mathematical models, and

4.b) development of scientific authoring tools for publications which exist only in the networked environment.

4.b.1) This will include integration of HTML+ WYSIWYG authorial and editorial tools, multidimensional data visualization applications, molecular modelling and database management tools into an interactive scientific publishing environment.

**System Diagram:**



- Definitions:**
- HTML+:** Hypertext Mark-up Language -- This is the language that World Wide Web databases are encoded in, and that Mosaic interprets.
  - IRV Server:** UCSF CKM's Interactive Remote Visualization Server -- This allows interactive real-time visualization tools to be embedded into Mosaic documents.
  - Vis:** UCSF CKM's distributed remote volume visualization tool
  - Midas:** UCSF CGL's molecular visualization package
  - SVM:** Sequence Visualization Module -- An as-yet unnamed tool for graphical display of genetic sequence data.
  - ITP:** Informal Text Phrase -- A user-entered search term, or a word or phrase of text that the user highlights from within a document.
  - URN:** Universal Resource Name -- A persistent, location-independent identifier for an object.
  - URL:** Universal Resource Location -- The address of an object. It contains enough information to identify a communications protocol and retrieve the object.
  - URC:** Universal Resource Characteristics -- Any combination of one or more URNs or URLs with meta information (e.g. author, format, compression method).

Rey

**Description:**

The system will draw from of a number of fundamental databases including bibliographic data (Medline) in the form of MARC records, journal publication data (Red Sage) in the form of SGML header and Postscript files, encyclopedic reference text data (CMDT) stored in an object-oriented SGML database, volumetric anatomical data (Visible Human Project) stored as NCSA HDF datasets, protein structure data (Protein Data Bank) stored as PDB files, and genetic sequence data (Genbank) stored as compressed ASCII strings (? , I'm guessing about Genbank).

These databases will reside behind a Z39.50 interface layer which yeilds, to the requesting client, the respective dataset in its native form. This data then goes through a translation layer where the data is either translated directly into HTML+ (Medline, Red Sage, CMDT) or loaded into a native-data visualization tool (Visible Human, PDB, Genbank). The HTML+ code is then passed to a set of HTML+ servers, which can be browsed by the Mosaic client. The visualization data is handled differently. The graphical I/O of the relevant visualization tool is passed to an interactive remote visualization (IRV) server, which handles both mapping of the display output from the visualization tool onto embedded live-visualization windows within the Mosaic-browsable HTML+ documents, as well as capture of user-entered mouse and keyboard events within the visualization windows and transmission of those mouse and keyboard events back to the relevant visualization tools. The user, browsing the system with the project's enhanced version of the Mosaic client, is presented with data and visualizations derived from these various databases, yet embedded into coherent, multimedia Mosaic documents.

For multimedia documents that have been explicitly pre-composed, the linking of these various data resources can take the form of universal resource names (URNs) that are encoded as tags into the HTML+ documents. This is passed to the system's semantic domain name server, for resolution of the information object's location and retrieval means. The URNs are used as indices in order to look up the relevant universal resource characteristics (URCs) in a URC database, which yeilds the universal resource location (URL), or physical adress, of the information object in question.

Semi-automatic means will be provided for a user to search for arbitrary information objects on the system by either keying in a search word or phrase, or by highlighting a not-already-hyperlinked section of text that (s)he happens to be viewing within the Mosaic client at the time. This informal text phrase (ITP) is then passed to the semantic domain name server, which passes it on to a universal resource thesaurus (which will incorporate elements of the NLM's UMLS system). The thesaurus compares the ITP to its database of terms and phrases and returns a rank-ordered list of URNs that are likely to match the object in question. These URNs are then passed to the URC database for resolution of URLs that point to information objects on the Internet that are most likely to match the ITP that the user employed to initiate the search. The user is presented with a rank ordered set of textual descriptions of likely matches which are hyperlinked, via their URLs, to the data in question. Clicking upon a selection from this list loads the related data into the relevant visualization server (IRV) or HTML+ server, and a second Mosaic window pops up to allow viewing or interaction with that dataset.

A set of authoring and editing tools will be designed to allow the interactive WYSIWYG creation of HTML+ documents, as well as allowing the embedding of visualizations, etc., which can be created using the interactive remote visualization tools, and which can use data from the various scientific databases mentioned above. Alternatively, the author can use his/her own datasets, which would be uploaded to an Internet-accessible World Wide Web server. The journal editor can use the same set of tools to edit submitted articles and to communicate changes to the text with the author. This, of course, would occur in a private, access-controlled, area of the system, so that confidentiality of the material to be published can be controlled.

Other private, access-controlled HTML+ servers will be used to administer the peer review process. A modification of NCSA's Mosaic-based group annotation server will be developed to allow the journal editor to exercise precise control and documentation of each reviewer's comments and suggestions.

**Contributions:**

## UCSF CKM:

- Development of Z39.50-compliant experimental (subset) databases for storage of Visible Human data, PDB data, and Genbank sequence data.
- Cooperation with AT&T in the development of an object-oriented SGML-based database for the Handbook of Current Medical Diagnosis and Treatment (CMDT)
- Development of an experimental Z39.50 interface to Medline data ( will be unnecessary if UC's DLA can provide such an interface to Melvyl Medline early enough into the project timeline)
- Development of translator servers to translate Medline MARC records, CMDT SGML data and Red Sage SGML/Postscript data into HTML+
- Development of a set of HTML+ documents that act as browsers to Medline, CMDT, and Red Sage
- Refinement and further development of Vis to allow better distribution of computation and better integration with Mosaic.
- Cooperation with CGL to adapt Midas for integration within Mosaic, and to identify and adapt a suitable program for graphical display of genetic sequence data.
- Refinement and further development of the interactive remote visualization server, and its incorporation (with NCSA's help) within the Mosaic environment.
- Development, in cooperation with NCSA, of an enhanced version of the Mosaic client to allow easier integration of external programs within Mosaic-readable documents.
- Development, in cooperation with Springer-Verlag and NCSA, of an interactive WYSIWYG editor for creation of HTML+ documents, and for embedding visualizations created using CKM's IRV tools, as well as development of a modified version of NCSA's group annotation server to support the peer review process.
- Development, in cooperation with AT&T, of an object-oriented SGML-based URC database
- Development, in cooperation with UCSF's CGL, UCSF's Radiology Dept., Washington Univ., and AT&T, of a Semantic domain name server and a URN Thesaurus, based upon AT&T's object-oriented SGML database technology.
- Development, in cooperation with UCSF's CGL, UCSF's Radiology Dept., Washington Univ., and Springer-Verlag of a set of sample content for use in evaluating the effectiveness of the system, as well as for demonstration of the results of the project.

## UCSF CGL:

- Cooperation with CKM to adapt Midas for integration within Mosaic, and to identify and adapt a suitable program for graphical display of genetic sequence data.
- Contributing to the refinement and further development of the interactive remote visualization server, and its incorporation (with NCSA's help) within the Mosaic environment.

- Development, in cooperation with UCSF's CKM, UCSF's Radiology Dept., Washington Univ., and AT&T, of a Semantic domain name server and a URN Thesaurus, based upon AT&T's object-oriented SGML database technology.
- Development, in cooperation with UCSF's CKM, UCSF's Radiology Dept., Washington Univ., and Springer-Verlag of a set of sample content for use in evaluating the effectiveness of the system, as well as for demonstration of the results of the project.

Washington University:

- Development, in cooperation with UCSF's CKM, UCSF's CGL, and AT&T, of a Semantic domain name server and a URN Thesaurus, based upon AT&T's object-oriented SGML database technology.
- Development, in cooperation with UCSF's CKM, and UCSF's CGL., and Springer Verlag of a set of sample content for use in evaluating the effectiveness of the system, as well as for demonstration of the results of the project.

AT&T Bell Laboratories:

- Development of Z39.50 interface to the RightPages server..
- Cooperation with CKM in the development of an object-oriented SGML-based database for the Handbook of Current Medical Diagnosis and Treatment (CMDT)
- Development, in cooperation with CKM, of an object-oriented SGML-based URC database
- Development, in cooperation with UCSF's CGL, UCSF's Radiology Dept., Washington Univ., and CKM, of a Semantic domain name server and a URN Thesaurus, based upon AT&T's object-oriented SGML database technology.

Springer-Verlag:

- Development, in cooperation with UCSF's CKM and NCSA, of an interactive WYSIWYG editor for creation of HTML+ documents, and for embedding visualizations created using CKM's IRV tools, as well as development of a modified version of NCSA's group annotation server to support the peer review process.
- Development, in cooperation with UCSF's CKM, and UCSF's CGL., and Washington Univ. of a set of sample content for use in evaluating the effectiveness of the system, as well as for demonstration of the results of the project.

NCSA:

- Cooperation with CKM in developing an enhanced version of Mosaic to allow easier integration of a client module for CKM's interactive remote visualization server.
- Cooperation with CKM and Springer-Verlag in the modification of NCSA's group annotation server to facilitate the peer-review process.

**Personnel:****Co-Investigators:****UCSF:**

Library & CKM: Richard Lucier, David Martin, Zoe Stavri, Ph.D., Cheong Ang, Marc Salomon

Radiology: Tom Budinger, Ph.D.

Molecular & developmental Biology: Tom Ferrin, Ph.D., Charles Ordahl, Ph.D.

**Washington University (molecular biology):** Toni Kazic, PhD

**Bell Laboratories:** Ed Szurkowski, Guy Story

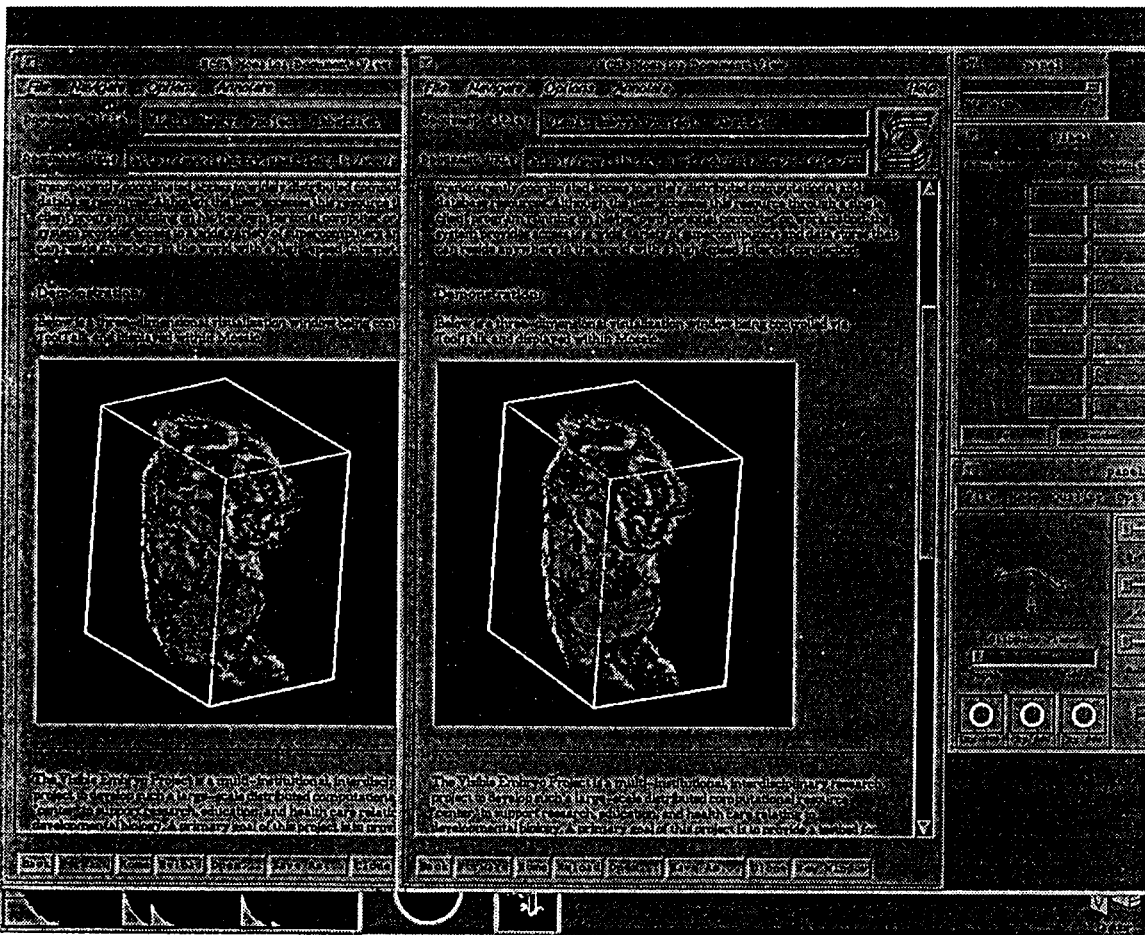
**Springer Verlag:** Bob Badger, PhD

**NCSA:** Joseph Hardin, PhD, & Mosaic development group

**SFSU:** Computer Science Dept. MS students

**Timetable:** 4 years

**Budget:** 1.2 \$M/year



**Figure 1:** A stereo-pair illustration of interactive real-time 3-dimensional human embryonic volume reconstructions embedded within an NCSA Mosaic document. This technology was developed by the Center for Knowledge Management at the University of California, San Francisco, and was demonstrated there in November, 1993.