I-DAM: An Active Digital Archive Management

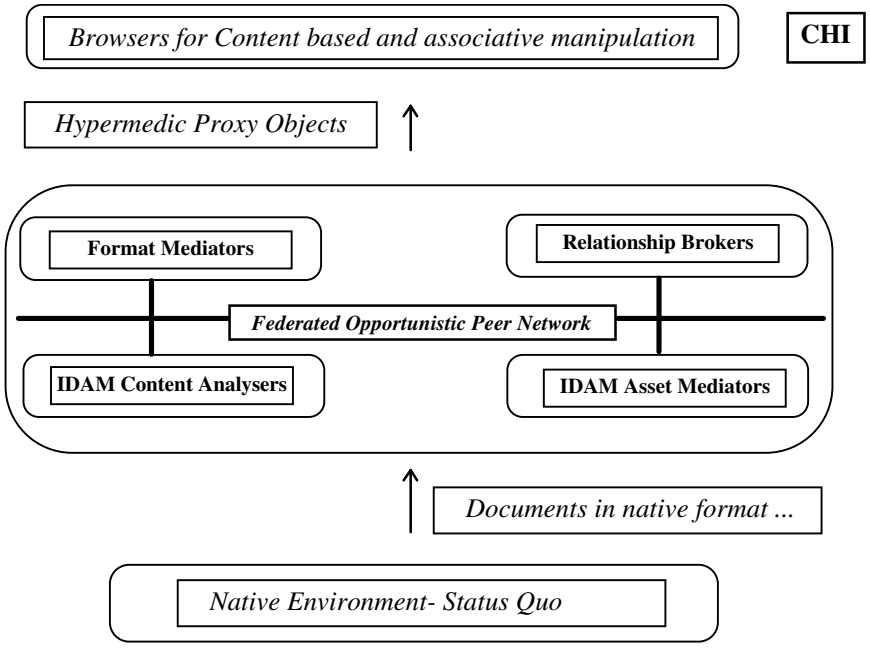
Raman Kannan
Center for Technology Transfer and Development
Monmouth University

Motivation

The power of information is not in how much there is but in how well it is harnessed and in this context the cyberspace as we know it, leaves much to be desired. Consider the internet today: (1) our workspace is inundated with information; (2) The "internet" (NII) facilitates the infrastructure; (3) MOMs, RPCs and ORBs facilitate connectivity and communication; and (4) the WWW technologies (HTTP, HTML and Browsers) address CHI factors in interfacing with these networked assets. And, yet, *the missing link*[1] is that these assets are accessible (1) if and only if they were available in HTML compliant format and (2) if and only if the precise locations were known and codified ahead of time. In other words, *the islands of automation* of the 80s have mutated into *islands of information* of incompatible formats. What we need is a comprehensive framework: (1) to leverage information on demand; (2) to render them accessible, regardless of the location and format in which they exist; and (3) to exploit the semantic relationships latent in these assets. Most importantly, we must do so not by creating another circle of protocols and tools but by customizing available technologies to bridge these worlds..

Info-Dam (I-DAM): The Metaphor, Architecture and Requirements (AKA Goals)

Info-dam -an active document archive management- is a framework to harness the true power of very large persistent information bases. *Info-dams* will be effective in controlling and managing information in the digital universe much as *dams* are effective in controlling and harnessing perennial rivers in the natural universe. As shown in the diagram below, Info-Dam is a collection of independent agents that cooperatively facilitate archival, retrieval and navigation, at the semantic level.



While there is no single tool like I-DAM, our strategy is to partition the IDAM task into smaller but distinct services and integrate (find or compose when necessary) cohesive tools that provide these services.

[1] No pun intended.

There are two major aspects to our proposal: (1) the IDAM framework to assemble a society of agents (brokers for mediation, transfer, transformation, CHI); and (2) a model for classifying the content, detecting relationships, and transforming documents into hypermedic objects. Furthermore, *content based and associative manipulation* is central to the IDAM model of interaction. To that end, whenever information assets are introduced into IDAM environment, they will be automatically analyzed for conceptual relationship using various techniques. Then using these relationships and a set of translation rules related assets will be gathered (or linked) and transformed into suitable formats. Finally, these transformed documents will be rendered using the WWW technologies at the mechanistic (transport and mechanistic) levels, on demand.

Besides the basic requirements (retrieval, search and navigation) we have identified three central requirements to our effort:

- Unobtrusiveness: Our main goal is to facilitate an environment where information assets exist in their natural format so that: (1) I-DAM users continue to create and manage documents in the usual manner and (2) existing information assets (legacy) are transparently incorporated into I-DAM environment.
- Universality: I-DAM effort must provide an acceptable level of service for *all* information assets. This is essential to counter (1) the explosive growth in the number of information processing tools; (2) associated formats; and (3) to breed a degree of familiarity in the individual user.
- No New Artifact: Leverage existing tools, and protocols, at all levels. This is one of economic necessity and moral obligation for those of us professing software engineering in grass-root academic centers.

Summary, Preliminary Results, and Concerns

In this position paper we have presented the motivation and a framework for incorporating any given document into a hypermedic environment. We posit that the success and acceptance of global digital library are predicated upon our ability to incorporate existing (and new) documents into WWW space. Furthermore, we believe that acceptable frameworks for digital libraries must preserve the original content and form of the document while rendering them in a form (HTML compliant) suitable for hypermedic interaction. Work is in progress to assemble a complete instance of the IDAM framework using readily available tools. We have devised a workable scheme by dividing the documents into structured and unstructured assets. For structured assets, such as High Level Programming Language components, we have implemented a prototype (for Language C, VHDL) using commonly available text processing tools such as tcl, perl and awk. For processing unstructured assets we are evaluating a hybrid approach incorporating various techniques (indexing engines at the lowest level, case based reasoning at the highest level). We are experimenting with Orbeline CORBA services to manage documents in a networked enterprise. End users browse the HTML counterparts using readily available Web browsers.

There are numerous unanswered questions and at the symposium we hope to initiate discussion, identify issues and requirements, resolve some of these questions, and refine our proposal. In particular:

- What are the appropriate interfaces for mediator, relationship broker and content analyzer services? Can we standardize? Are KQML and KIF relevant?
- What are the appropriate protocols for dynamic binding between web clients and situated ORB agents? Are KQML and KIF relevant?
- How to extend ORB IDL and HTML to declare/specify content semantics? Are KQML and KIF relevant?
- How to synchronize proxy objects and the underlying base documents?

A research stipend from ASSET Corporation, Holmdel, NJ, an equipment subsidy from CTTD, and graduate research stipends from the Graduate School are acknowledged. Supporting bibliography may be obtained from the author.