# Paradigms of Information Access In the 21st Century

John Mallery
M.I.T. Artificial Intelligence Laboratory

This seminar outlines some of the challenges and presents current thinking in an MIT AI Lab research project which grew out of an experiment during the 1992 Presidential Elections. The project has two main tracks:

- **Intelligent Servers:** An email server substrate is under development to service various needs of a complex organization. A prototype electronic publication system for White House press releases aims to automatically categorize documents and route them to interested publics. A knowledge-based system processes the inevitable backwash which arises when email is so widely disseminated. An adaptive survey research facility has been developed to chart the dissemination pattern of White House releases and discover how they are used. An object-oriented data exploration environment uses generic protocols to connect machine learning algorithms with realworld data, for instance, survey returns.
- **Natural Language:** A more ambitious research track aims to develop natural language understanding systems which can syntactically parse and semantically represent free text not just to more effectively deliver/retrieve information to relevant people but also to serve as intelligent guides in the infoswamp. As an early prototype, we are developing an intelligent receptionist that answers personnel questions via email and directs inquiries to the appropriate person in an organization. Later systems will conduct more extended dialogues with people to ascertain their goals and connect them with the appropriate information or service.

As a bridge between the workable and the AI-complete, we consider a variety of intermediate strategies that can help large groups of people and organizations communicate more effectively as they cooperatively solve problems or carry out complex tasks. Distributed hypermedia must figure prominently in our intermediate strategy, not just because of its phenomenal growth rate, but also, because these distributed semantic networks are the primitive forerunners of a worldwide knowledge base. In this revolution, it is increasing apparent that a strong convergence exists between digital networks, information, and knowledge

## Menu

1. Setting
   - The Digital Network Revolution
   - Managing Complexity
   - Challenges for Wide-Area Communications
2. Approach
   - Current Research
   - Research Strategy
   - Technologies
   - Paradigms of Information Routing
3. Projects
   - Email Server Substrate
   - White House Publications System
   - Pollster's Assistant
   - Data Exploration Environment
   - Intelligent Receptionist
   - Correspondence System
   - Transaction Stereotypes
   - Natural Language Processing
4. Conclusions
   - From WWW to WWKR
   - Consequences

## The Digital Network Revolution

- Over 40 million people worldwide have access to email.
- The number of internet hosts is growing at 17 percent per month, and there are already over 2 million connected.
- 8 terabytes of data cross the US internet every month, whereas the figure is 1 teraflop for Europe.
- Telecommunications giants and computer companies are positioning for what many believe has become a single multimedia market.
- The White House is publishing press releases and major documents electronically and also accepting email to the President and Vice-President.
- The US Congress is experimenting with email and other internet connectivity.
- The OMB A-130 Circular directs Federal agencies to devise plans for interfacing electronically with the public and each other.
- Legislation in support of the National Information Infrastructure is moving through Congress, but aims largely at funding research and pilot projects to determine where and how Federal funds should be spent.
- Nobody is quite sure what the outcome will or should be, nevertheless none want to be left behind.

## Managing Complexity

- **Uniform Access:** Too many different modes of information access require too much effort.
- **External Indexation:** Inability to locate new information relevant to one's purposes.
- **Internal Indexation:** Inability to recall previously acquired information.
- **Understanding Support:** Inability to determine which information is relevant and how much of is needed.
- **Integration:** Inability to distill and digest all relevant information.
- **Decision Support:** Inability to relate information to goals.
- **Realtime Updating:** Inability to adjust to changing information.
- **Computer Support:** Insufficient time and human resources to process relevant information.

## Challenges for Wide-Area Communications

- Complete distribution of focused information.
- Focused document retrieval.
- Distributed data-integration and analysis.
- Automatic management of information bottlenecks.
- Enabling large-scale collaborative and interactive problem-solving.
- Distributed (inconsistent!) knowledge representation and cooperative computing.
- Integration of organization design with computational infrastructure.

## Current Research

1. **Artificial Intelligence**
    - Email Server Substrate - Mallery
    - Knowledge-based Postmaster - Shrobe
    - White House Publications System - Mallery, Hurwitz, Shrobe
    - Pollster's Assistant - Hurwitz, Renaud, Mallery
    - Data Exploration Environment - Renaud, Mallery
    - Intelligent Receptionist - Katz
    - Correspondence System - Davis, everybody
    - Transactional Stereotypes - Hurwitz, Malone, Mallery
    - Natural Language Processing - Berwick, Katz, de Marken, Mallery
2. **Social Science**
    - Political Communication - Hurwitz
    - The Digital Organization - Tom Malone (MIT Sloan School)
    - Sociology of Communication - Lee Sproull (Boston University)

## Technologies

1. **Network Transport Protocols:**
    (a) Store and Forward Mail via SMTP
    (b) News Propagation via NNTP
    (c) Document Retrieval via Z49.50
    (d) Hypertext Transfer via HTTP
    (e) Other (*e.g.* FTP, Gopher)
2. **Varieties of Servers:**
    (a) Email Servers
    (b) Hypermedia Servers
    (c) *Knowledge-based Servers*
3. **Methods to Structure Communications**
    (a) Subject Line Commands (remote procedure call?)
    (b) Automatic Form Processing
    (c) Multimedia User interfaces
    (d) Statistical Document Categorization
    (e) Sparse Syntactic Parsing
    (f) Syntactic Parsing with Surface Semantic Representation
        - Approximate Representation
        - *Accurate Surface Semantic Representation*
        - *Deep, Knowledge-Rich Understanding*

## Research Strategy

- **A two track approach:**
  1. Pushing workable technologies (*e.g.*, email servers, form processing) in operational systems.
  2. Working on difficult technologies (*e.g.*, NLP, machine intelligence).
  3. When new techniques prove effective, carry them over to the operational systems.
- Operational systems provide realworld constraints and *data* that focus research on the important issues.
- Ambitious research explores new approaches, some of which may be transferable.
- Protocols at key system interfaces make plug-compatible testbeds possible for:
  1. Statistical document categorization
  2. Machine learning (feature vectors)
  3. Syntactic parsing/generating
  4. Knowledge representation
  5. Inference/Learning
- Different approaches can be tried and the best one used in the operational system.
- Bottom-up methods with high generality are preferred because they lay a solid foundation to build on. (Don't run before you can walk).

## Paradigms of Information Access

- **Information Source:** Document base, person, program, knowledge base.
- **Information Sink:** Person, program.
- **Routing:** Connecting the source with the sink is only half the story.
- **Purpose:** It requires intelligence to be a helpful information source because the purposes of the information user need to be considered in order to increase relevance.
- **Felicitous Response:** Understanding the goals of the information seeker can allow a system to carry out more of the information processing task.

## Paradigms of Information Routing

1. **Static Routing:**
   - **Mailing Lists:** fixed pointers that change slowly.
   - **Hierarchical Mailing Lists:** Taxonomically arranged mailing lists.
2. **Dynamic Routing:**
   - **Taxonomic:** Routing on the basis of boolean combinations of keywords associated with categories in a multidimensional taxonomy (Top-down)
   - **Filtering:** Routing on the basis of user defined statistical filters on messages. (Bottom-up)
   - **Mixed-Strategy:** Routing combines methods taxonomic and filtering approaches with some syntactic parsing (*e.g.*, noun phrase parsing) and possibly some limited knowledge (*e.g.* sense disambiguation).
3. **Knowledge-based Routing:**
   - **Semantics-based:** Routing based on syntactically parsing and semantically representing message so that recognizers at the semantic level can assign relevance to parts of messages.
   - **Understanding-based:** AI-complete system does not make errors in representing the message so that it can do the right thing.

## Email Server Substrate

- Subject line command processing.
- Form processing based on CLIM.
- Reliable task execution.
- Persistent object database.
- Myriad issues concerning networks and mailers.
- *Specification/Protocol for a remote GUI for forms.*
- *GUI for form definition and release.*
- *Scaling for high volume.*

**White House Publications System**

1. Automatic Subscription Maintenance.
2. Taxonomic Document Routing.
   - Top-down by boolean combinations of categories.
   - *Bottom-up by words and phrases.*
3. Document Retrieval.
   - Top-down by boolean combination of categories.
   - Bottom-up by WAIS-style queries.
4. Standing Usage Survey.
5. Publications Interface.
6. Failed Mail Processing.
7. *Automatic Document Categorization.*

## Pollster's Assistant

- Survey research building on automatic form processing.
- Hierarchical and adaptive questions.
- User friendly follow-up to failing or omitted queries.
- Portable Interface for rapid survey design, release, and analysis.
- Machine learning algorithms to explore returns.
- *Acquiring relational models of people's beliefs and norms.*

## Data Exploration Environment

- The Feature Vector Editor.
- Protocols mediate interactions between the user, data, and analysis techniques.
- Rule induction from non-rectangular data.
- *Integration with email-based automatic survey research tools.*
- *Interface for defining and combining dataset codebooks.*
- *Generalization to multiple, remote datasets: WWW for data.*
- *Integration of additional machine learning algorithms.*

**Intelligent Receptionist**

- Natural language email server (subsumes form-based servers).
- Answers questions based on a knowledge representation.
- Finds larger pieces of text or tables that answer questions.
- Serves information about the MIT AI Lab.
- *Might refer inquiries in the Media Affairs Office.*

## Correspondence System

1. **Robomail:** Returns form letter responses from a stock of predefined responses.
    - **Statistics:** Use statistical methods to match documents to replies. (Problem with negation.)
    - **Form Processing:** Use forms to discriminate and return the correct response.
    - *Generation: Generate custom responses from a knowledge base of answers (the representation of the response forms).*
    - *NLP: Parse and represent inbound correspondence to determine the appropriate response form.*
2. **Service Processing:** Use a form-processing email server to act as an self-service information desk, doling out information on request and providing a range of services.
3. **Opinion Sampling:** Use automatic surveys to elicit more than whether someone is "for or against" an issue or policy.
4. **Argument Hypertext:** Allow people to collectively construct hypertext reasons *for* or *against* issues or policies. People could also endorse existing links.
5. **Reason Sampling:** Use form processing to acquire relational models of people's opinions.

## Transaction Stereotypes

- Studying the interactions between people and organizations can uncover transactional stereotypes that can set expectations for people and guide form processing email servers.
- A standard library of transaction stereotypes can make it easy to set up an electronic interface to the public and provide consistency across organizations.
- Representing the goals for which a service is appropriate can allow a form-processing email server to minimize the number of queries required for people to get what they want.
- Similarly, a natural language system can ask questions of a person in order to disambiguate what they want, when their intention is only implicit in their message.

## Natural Language Processing

- **Syntax**

  1. A free text parser that handles errors gracefully.
  2. A declarative parser/generator pair that correctly analyzes text and regenerates sentences.
  3. A multi-lingual, principles and parameters parser/generator pair that relies on a small number of principles and can be easily adapted to new languages.

- **Lexicon:** A general-purpose lexicon needs to support the various syntax efforts.

- **Semantics**

  1. A interface protocol between syntax and semantics, allowing syntax modules to be swapped in and out.
  2. A knowledge representation that makes no epistemological precommitments for researchers who might use it.
  3. A reference system that can reliably find intersentential coreferents.
  4. A lexical categorization system that can recognize semantic configurations.
  5. A constraint-based graph walker that can drive a sentence generator as well as build lexical categorizers.

## Previous Work at the AI Lab

- Email servers in the 1992 Presidential Campaign.
- Various machine learning algorithms.
- Winston induction and analogy work.
- Katz START Natural Language System.
- Duffy-Mallery Relatus Natural Language System.
- De Marken Sparse Parser.
- Fong Principles and Parameters Parser.

## From WWW to WWKR

- A hypermedia system is a semantic network whose nodes contain opaque boxes of text, audio, or video which the system does represent or understand.
- As distributed hypermedia systems grow in complexity and sophistication, increasing amounts of information about these opaque boxes will be encoded in the link structures, and this will allow systems to respond more flexibly.
- As more knowledge becomes encoded in the link structures, knowledge will become more readily accessible and ever greater opportunities for cooperative computing will emerge.
- As this evolution unfolds, the link structure will be increasingly used to constrain search over the worlds networked computers.

## Consequences

- How can we make complex systems accessible to everyday people?
- How can organizations become more effective through better information processing and intelligent agents?
- How will tighter information integration change organizational dynamics? Heightened conflict or oppressive homogenization?
- What will be the consequences for political systems, especially democratic ones such as our own?