

# **The Digital Libraries Initiative**

## **A USA Federal Program of Research and Applications**

Stephen M. Griffin  
Program Director, Digital Libraries Initiative  
National Science Foundation  
USA  
sgriffin@nsf.gov

### **Abstract**

This presentation will give an historic overview of the US Digital Libraries Initiative (DLI) and highlight current project activities. The first part of the talk will note developments, technological and social, which have helped create a modest global information infrastructure over the past decade and speculate on future directions. The effects of these developments on social, institutional and scientific practices will also be noted. The second part of the talk will discuss individual DLI project activities.

### **Key Words:**

Digital Library, Digital Libraries Initiative, Global Information Infrastructure, International Collaboration

### **Introduction**

Since the announcement of “Research on Digital Libraries Initiative” in 1994, research and technology integration efforts in this area have proliferated. The label “digital libraries” has become a powerful metaphor; one that has inspired innovative thinking about environments in which people use geographically distributed information, computing and communications systems. Over the same period, the definition of “digital libraries” has continued to evolve, technologies have advanced, and creation of digital content has multiplied dramatically.

A future in which vast amounts of digital information will be easily accessible to large segments of the world’s population is seemingly within reach. The pace of technological innovation and the rapid emergence of powerful, new information infrastructure components has altered the way individuals, communities of all sizes, and institutions perceive and conduct their day-to-day affairs. We are experiencing a time of dramatic and far-reaching change, with implications that we are unable to predict, or even imagine. Past experience offers little to assist us in anticipating the future.

While we do not yet have a full understanding of the causes of the rapid transition toward a global information environment, certain factors can be credited with great certainty. Among them are:

- The continuing development of commodity high performance hardware and software which is transforming the availability, capabilities, and cost of computing and network access for many people. This has led to low cost, powerful computing devices connected to a high bandwidth global network infrastructure.

- The spread of the Internet hosts and connections worldwide, and the nearly universal adoption of the World Wide Web (WWW) as a primary access framework. This has led to large segments of society valuing and information environments that allow individuals to contribute and the world to share information of all types. (The WWW, even with its primitive capabilities as compared to the environments that exist in some organizations, is still the primary source of reference materials for many users throughout the world.)
- Dramatic advances in the means to create new digital content and to convert information of all types originally recorded on other mediums into digital form
- Growing consensus on common approaches, frameworks, formats, protocols, etc., to describe and structure information created by different people in vastly different circumstances, in order to achieve interoperability between the systems that operate on the information.

Although the results of our current efforts may not be fully known or appreciated for decades, we will be judged by those who use the things we build now, on how quickly we perceived and took advantage of opportunities presented, how wisely we selected between alternative paths, and most of all, whether we allowed local and parochial interests to interfere with creation of resources for the common, global, social good.

Digital libraries are meant to provide *intellectual access* to large, geographically distributed stores of information of all types. This implies incorporating semantic capabilities into information environments which significantly advance access beyond *electronic access* to raw data - the bits. Digital libraries research is concerned with developing concepts, technologies and tools to gain access to the fuller knowledge and meaning contained in digital collections. As examples, for users this means intelligent search, retrieval and presentation tools and interfaces; for content and collections providers this means new information types, structures, document encoding and metadata for enhancing context; for system builders this means designing hardware and software systems capable of interpreting users' requests, federating collections and selecting from a vast multitude of possibilities to provide what is *desired and needed* - not merely what is requested as bounded by a particular users knowledge and imagination. A primary challenge for funding agencies is to discern a reasonable balance between resources directed to content structure and markup (metadata) at the early stages of creating digital libraries or basic research aimed at building more intelligent software to compensate for lack of this later on.

A major goal of DLI-Phase 2 is to foster advances with the potential of measurable impact on research, education and commerce in an increasingly complex global environment. The current state of the world's store of digital content is one of explosively increasing amounts of information created by many people or data gathering instruments, in many forms, stored in many formats on millions of systems located around the world, increasingly interconnected via electronic networks. Different pieces of this elaborate complex are managed by different individuals and organizations, embracing different values and principles, and speaking different languages.

Digital libraries are designed to enable more people to better create and use vast amounts of distributed information in relatively uncontrolled, dynamic, open environments. In such a rapidly evolving environment, traditional roles and distinctions between content creators, providers and consumers become blurred, breakdown and even reverse. Because digital libraries are dynamic and widespread, with content, content organization, content location, delivery systems and users changing frequently and instantaneously, they require new thinking and models for information management, access, use, and long-term

archiving and preservation. Experience has shown that digital technologies are pushed by application to new domains - both scientific and non-scientific. Involvement in new subject areas informs technology research. Of particular note in the DLI - Phase 2 program are interdisciplinary efforts in the natural sciences, humanities and cultural heritage applications areas.

International cooperation and collaboration is key to attaining the goals outlined above. Global coordination in the way digital libraries are developed and used is essential to accessing globally distributed, multilingual information and enabling users to easily access digital collections, regardless of location, language or formats, for research, education and commerce and other purposes. Only through joint international research and application efforts can world-wide systems for accessing information sources on the internet be realized.

The development of a global information environment requires international cooperation and collaboration in many domains. At the lowest level, it requires joint work and agreement on *interoperable technologies* to enable creation and common use of many kinds of information. It requires joint work and collaboration on *content development* to help create and make available on the internet useful and usable information of cultural, social and scientific value. It requires joint work on the *development of standards* for ensuring consistency and long-term sustainability of resources which are geographically distributed and independently administered. It requires joint work and collaboration on *protecting intellectual properties* in an open, fluid, global marketplace.

The Digital Libraries Initiative - Phase 2 is committed to increasing international collaboration in these areas. The following WWW addresses point to the program announcements, sponsored projects, and activities which have been funded to date. We look forward to a continually expanding program of support for joint international activities.

<http://www.dli2.nsf.gov/>  
<http://www.dli2.nsf.gov/intl.html>  
<http://www.dli2.nsf.gov/projects.html>  
<http://www.dli2.nsf.gov/intl.html>  
<http://www.dli2.nsf.gov/workgroups.html>

## A Presentation to the 18th Digital Library Workshop (DLW18)

### The DIGITAL LIBRARIES INITIATIVE A USA Federal Program of Research and Applications



www.dli2.nsf.gov

Stephen M. Griffin  
National Science Foundation  
sgriffin@nsf.gov

## Digital Libraries Initiative (DLI) Phase 1: Program Profile



- Sponsored by NSF, DARPA, NASA
- 1994 - 1998
- Six university-led projects; similar project model for each
- \$24M total over five years, ending fall 1998.
- A program of fundamental digital libraries research, testbed building and partnerships

DLW18

## Digital Libraries Initiative (DLI) Phase 1: Project/Research Focus



**Carnegie Mellon University: Digital Video Libraries**

- speech, image and natural language technologies integration

**Univ of Michigan: Intelligent Agent Architectures**

- software agents; resource federation; artificial service market economies; educational impact

**Stanford Univ: Uniform Access**

- interoperability; protocols & standards; distributed object architectures; interface design for distributed information retrieval

**Univ of California, Santa Barbara: Geographic Information Systems**

- spatially-indexed data; content-based retrieval; image-compression; metadata

**Univ of Illinois: Intelligent Search and the Net**

- large-scale information retrieval across knowledge domains; semantic search; SGML; user/usage studies

**Univ of California, Berkeley: Media Integration and Access**

- new models of "documents"; natural language processing; content-based image retrieval; innovative interface design

DLW18

## DLI Phase 1 Collaboration and Partnering

-- DLI Lead Institutions --  
Carnegie Mellon University    University of California, Berkeley    University of Illinois  
Stanford University    University of Michigan    Univ of California, Santa Barbara

Flow of Resources, Technologies, Knowledge, Intellectual Products

Computer & Communications	Publishers/Content Providers	Other Universities	Primary & Secondary Schools
CompuLink Digital Equipment Corp Xerox Corp Xerox PARC Intel Corp Apple Corporation Bellcore Eastman Kodak Co IBM Lockheed Interconnect Tech Corp Enterprise Integration (EIT) Berkore Interval Microsoft Corp Bell Atlantic Network Services AT&T Hewlett Packard United Technologies Softquad PARADIGMware Syrinx Hitachi	Elsevier Science Group Encyclopedia Britannica McGraw-Hill Publishers Dialog Information Services O'Reilly Wiley QED Communications John Wiley & Sons U.S. News & World Report MAT Publishing Tribune Company UNL  Professional Societies American Math Society (AMS) ACM IEEE American Institute of Aeronautics and Astronautics (AIAA) American Physical Society American Institute of Physics NCSA Association of Research Libraries	SIU Buffalo Univ of Miami Univ of Arizona Open University, U.K. Univ of Wisconsin Univ of Colorado MIT Cornell Univ  Libraries Project Safe Univ Libs USGS Library Library of Congress California State Library Sonoma County Library St. Louis Public Library New York Public Libs International Orgs ERIC	Project-local comm schools Fairfax County Public Schools Winchester-Thornton School Ann Arbor Public Schools Stuyvesant High School, NYC Shasta County Off. of Ed.  Government Agencies and Labs DMA, CTO US Navy USGS NASA, ARI Res. Agency of California San Diego Arm. of Govt.  Other/Non-Profits CNRI Environmental Systems Research Mellon Foundation Kellogg Foundation Gentry Foundation

## Digital Libraries Initiative - Phase 2



Core Sponsors: NSF, DARPA, NLM, LoC, NASA, NEH  
~\$R-10 million/yr for 5 years (beginning FY98)

- sponsor a full-spectrum of activities: fundamental research, content & collections development, domain applications, testbeds, operational environments, new resources for education and preserving America's cultural heritage
- address topics over entire DL lifecycle: information creation, dissemination, access, use, preservation, impact, contexts
- implement a modular, open program structure: add new sponsors, performers, projects at any time

### Program Goals:

new DL research, technologies and applications to advance the use of distributed, networked information of all types around the nation and the world

DLW18

## DLI Phase 2 Collaboration and Partnering

-- DLI2 Academic Institutions --

Flow of Resources, Technologies, Knowledge, Intellectual Products

Computer & Communications	Government Agencies and Labs	Professional Societies
CompuLink Digital Equipment Corp Xerox Corp Xerox PARC Intel Corp Apple Corporation IBM SRG International Oracle GE Interval Microsoft Corp Bell Atlantic Network Services AT&T Lucent Technologies Hewlett Packard Infomix Shup NEC Hitachi Sun Microsystems Hewlett-Packard Welch Allyn	Smithsonian Institution US Navy Los Alamos National Laboratory National Park Service California Academy of Sciences CA Env. Res. Eval. Sys. (CERES) CA Dept. of Water Resources San Diego Supercomputer Center USGS NASA/ARC Resources Agency of California S. California Earthquake Center Consortium of Research Libraries-UK UK Office for Library & Information Networking  Libraries/Museums Library of Congress California Digital Library New York Public Library NASA Ames Library USGS Library Museum Fine Arts, Boston	Modern Language Association ACM Oral History Association NCSA Association of Research Libraries Chicago Historical Society  Other/Non-Profits Mellon Foundation Patrick Humanities Institute Gentry Foundation Columbia Presbyterian Medical Center  International Orgs ERIC JISC DFU  Content Providers CNM The News Hour with Lehrer Dialog Information Services

## DLI2 Collaboration and Partnering Academic Projects Partners

University of Arizona	University of Liverpool
University of Bath	University of Maryland
University of Bristol	University of Massachusetts
University of California at Berkeley	University of Michigan
University of California at Davis	Michigan State University
University of California at Los Angeles	University of North Carolina
University of California at Santa Barbara	Old Dominion University
Carnegie Mellon	Oregon Health Sciences University
Columbia University	Oregon Graduate Institute of Science and Technology
Cornell University	University of Pennsylvania
Eickel College	University of Texas at Austin
Georgia State University	University of South Carolina
Harvard University	Southampton University
University of Illinois at Chicago	Stanford University
Indiana University	Swarthmore College
John Hopkins University	Tufts University
University of Kentucky	University of Washington
King's College, London	University of Wisconsin at Madison
University of Leeds	

DLW18

## Comparison of DLI with DLI - Phase 2

### DLI- Phase 1 (1994-1998)

**research**  
broad, technology-centered

**testbeds**  
for technology research

**content/collections**  
donated to projects

**infrastructure**  
limited testbed development

**context**  
primarily user evaluation

1994

### DLI - Phase 2 (1998-2002)

**research**  
refined technical scope; extend to new areas and dimensions in the DLI information lifecycle

**testbeds**  
for DL research with added emphasis on interoperability & technology integration

**content/collections**  
increased emphasis on content, collections development and management

**infrastructure**  
operational DLs with collections of value to domain and other "communities" of users

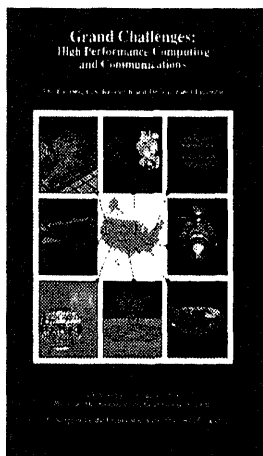
**context**  
understanding DLs in domain, economic, social, international contexts

1998

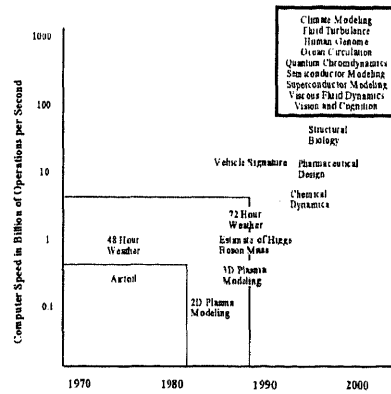
DLW18

## The Federal High Performance Computing and Communications Program

1992-1996

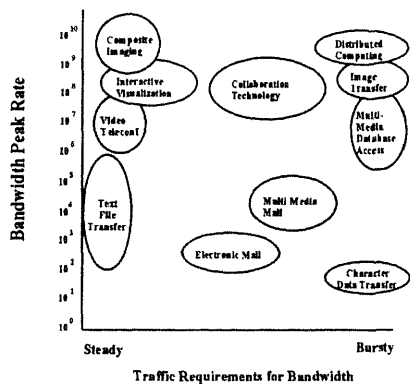


## Grand Challenge Requirements



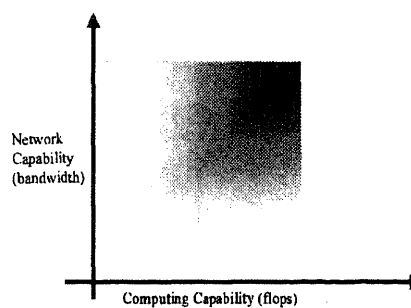
DLW18

## NREN Applications by Bandwidth and Traffic Characteristics



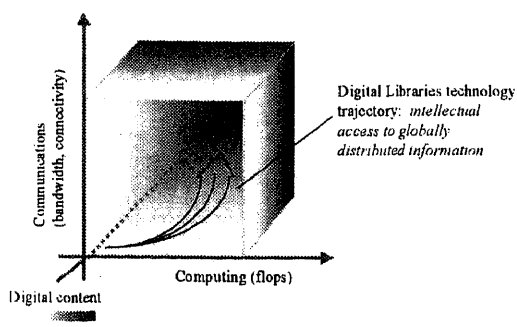
DLW18

## Two Dimensional Thinking of Early 1990s...



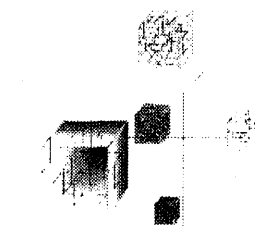
DLW18

### Three Dimensional Thinking of mid-90s...



DLW18

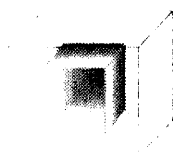
### Next: Advanced Functional Capabilities, Wide Use



Intelligent distributed knowledge environments present vastly different content at the user level yet maintain striking similarities at the digital level...

DLW18

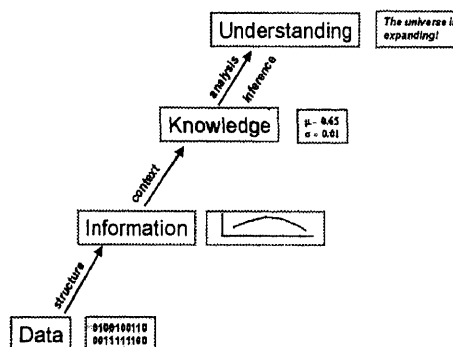
### Add Context and Structure to Digital Content in Early Stages of Preparation



experience has shown that adding metadata to digital content early on makes a digital library much more useful and inexpensive than trying to create more intelligent software to compensate for it later

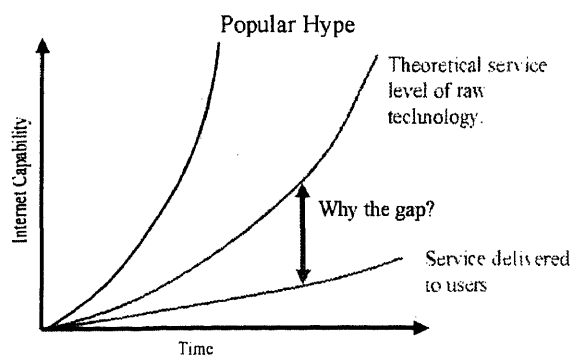
DLW18

### Evolution of Understanding in a Distributed Knowledge Environment



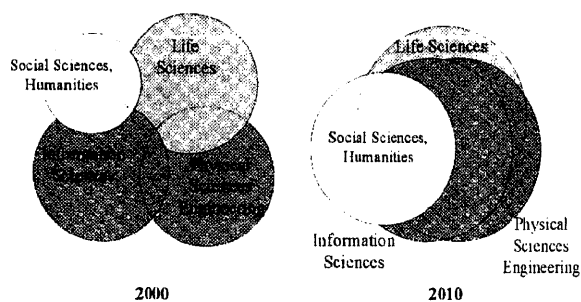
DLW18

### The state of affairs today ...



DLW18

### A Vision of Disciplinarity: The World in 2010



DLW18

Information technology is pushed by research and applications in other disciplines.

Computer Science is stressed by and enlivened by engagement in new topical problem areas.

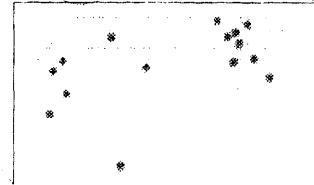
Interdisciplinarity beyond the sciences has much to offer.

DLW18

## Changing Scales and Contexts of Interaction and Collaboration



NSFNet StarTap Connections



DLW18

## International Collaboration

International collaborative efforts are essential to achieving a content-rich, balanced Global Information Infrastructure. Issues must be addressed through collaborations at many levels.

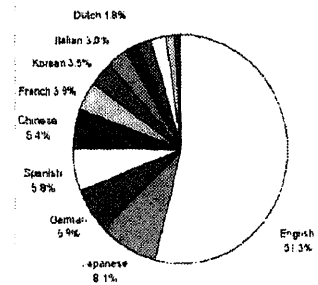
DLW18

## Languages and the Internet

April 1999

English	107.2M	56.5%
non-English	82.7M	43.5%
European	54.9M	30.0%

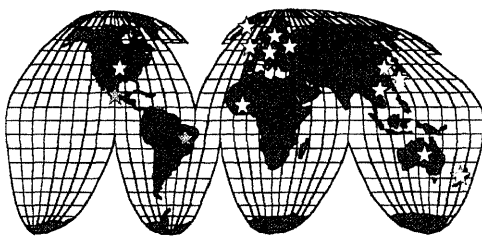
Online Language Populations (June, 2000)



<http://www.mcmanis.com/globetalk>

DLW18

## International Digital Libraries Collaborative Research Program



- Activities
- ~20 small awards
  - ~30 countries
  - Formal Programs with I/K/AISC, Germany/DFG, EU/EC

DLW18

## Making Global Digital Libraries Infrastructure Means

- ◆ Merging intellectual perspectives
- ◆ Dealing with heterogeneity at many levels
- ◆ Achieving interoperability at many levels
- ◆ Integrating information technologies
- ◆ Building large collections of great diversity
- ◆ Supporting functions beyond search and query
- ◆ New conceptualizations of the future (imagination)
- ◆ Global participation
- ◆ Economic and IP models for new information use
- ◆ ...

DLW18

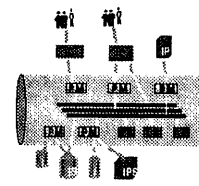
### Building Large-Scale Operational Systems

Our understanding of the impacts of digital libraries on social institutions and practices is limited because we do not yet have large-scale systems being heavily used to observe and analyze.

The reflexive behaviors of systems, supporting infrastructures and user populations become apparent when millions of people use digital libraries, not thousands.

DLW18

### Making Digital Libraries Infrastructure Requires Dealing with heterogeneity at Many Levels\*



PM: Protocol Machine  
L.S.: Library Service  
IC: Interface Client  
S.S.: System Service  
IPS: Information Processing Service

Stanford InfoBus CORBA distributed object technology

\* objects, collections, services, platforms....

DLW18

### Making Digital Libraries Infrastructure Requires Merging Intellectual Perspectives

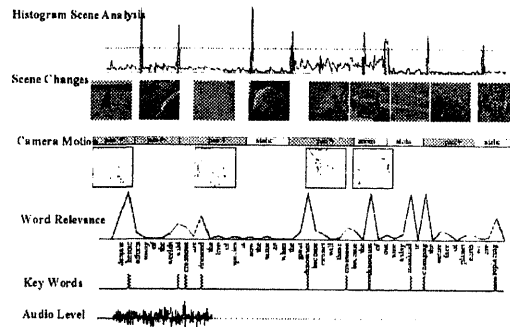
**Traditional Libraries Stress:**  
Service  
Selection, Organization, Structure for Access  
Centralization, Standards  
Physical objects & standard genres

**Contemporary Technological Capabilities (e.g. WWW) Stress:**  
Flexibility, Openness  
Rapid Evolution  
Decentralization (geographic, administrative)  
Digital objects, old and new genres

**Design Considerations for Digital Libraries & Beyond**

DLW18

### Making Digital Libraries Infrastructure Requires Application of Integrated Technologies



© 1997-1998 by UC Berkeley

DLW18

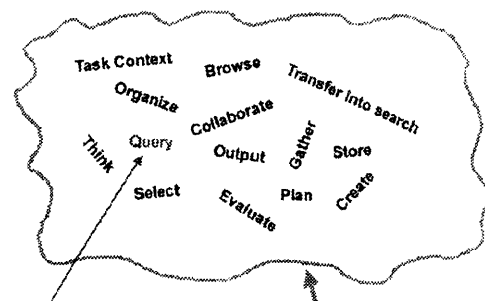
### Making Digital Libraries Infrastructure Requires Building Large Collections of Diverse Information

UC Berkeley Testbed Content

Type	Examples	Sep 99	
Documents	articles, EIRs, water reports	280,237 pp	67 GB
Images	DWR Cal Flora Corel Animals, etc	17,601 20,286 39,100 1,875	
	Total	78,862	474GB
Aerial photos	Suisun Marsh Sac-SJ Delta	1074 img	3.4GB
Sensor Data	Delta fish flow	30 days	02MB
GIS Data	dams, fish, watersheds, etc.	various	50MB
DOQs	SF Bay Area	219 img	33GB
DRGs	California		26GB
CalFlora DB	Occurrences	674,814	539MB
Other tables	MVD, streets	1,285,096	185GB
<b>Total</b>			<b>789GB</b>

DLW18

### Making Digital Libraries Infrastructure Means Supporting More than Query



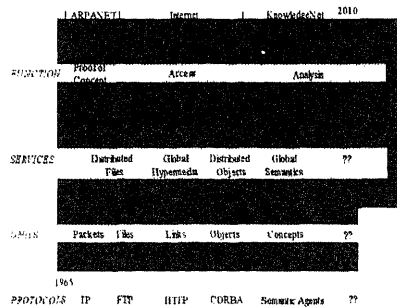
Today's Technology Centered Systems

User and Usage Centered

DLW18



## Making Digital Libraries Infrastructure Requires New Conceptualizations of the Future (imagination)



DLW18

## Basic Representations Of Music & Audio

Audio (e.g., CD, MP3)

Time-stamped Events  
(e.g., MIDI file)

Music Notation

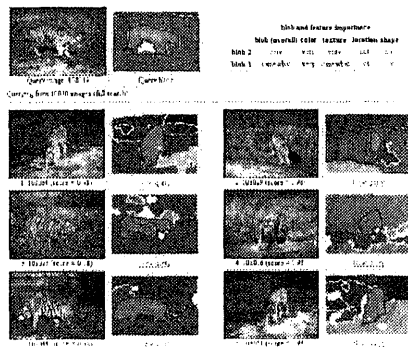


David R. Shaw  
C.I.R. Technical Information

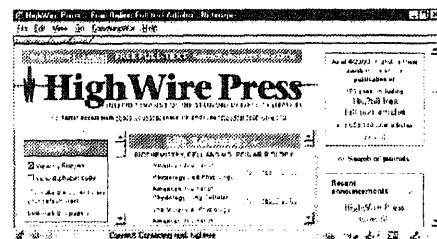
First International Symposium on  
Music Information Retrieval  
<http://cir.cs.umass.edu/music2000>

DLW18

## Blobworld Update



DLW18



DLW18

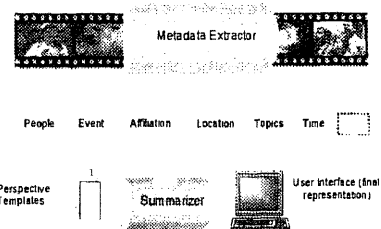
## Open Archives Metadata Set

Facilitate coarse granularity discovery  
Compatible with Dublin Core

Title	Author and Affiliation	Entry Date
Discovery Date	Subject	Abstract
Unique ID	Access URL	Comment

DLW18

## Informedia-II: Integrated Video Information Extraction and Synthesis



DLW18

### Example of a Large Data Object: Michelangelo's David



#### Laser scanning, digitization and computational rendering of Michelangelo's David

The scanning gantry is 7.5 meters high

480 individually aimed scans were performed. The resulting 3D raw data set exceeds 250GB. The digital model, accurate to 25mm, contains 2 billion polygons and 7 thousand color images. Laser scanning, digitization and computational rendering of Michelangelo's David.



David's left eye

By Professor Marc Levoy of Stanford University

DLW18

### Multilingual Information Access (MLIA) Systems

capabilities of a successful Multilingual Information Access (MLIA) system

- Accept a query in any language
- Return relevant documents in any language
- Present the gist or summary to users in any language

Summary Report of the NSF-EU Working Groups

Brussels, October 12, 1998

### Resource Discovery

#### Long-Term Challenges

- Formal characterization of user behaviors, roles, and expectations
- Examination of organizational models of federated information spaces and their impact on distributed resource discovery techniques
- Investigation of new document models (active, distributed, compound, etc.) and their impact on research discovery
- Development of integrity, reliability, and consistency models in decentralized distributed environment

Summary Report of the NSF-EU Working Groups

Brussels, October 12, 1998

### Goals for the Future

- Gather information and build collections  
(to better use what we have and discover what is missing...)
- Create new global communities  
(to communicate and collaborate)
- Make technology disappear  
(from our awareness and experience)

### For More Information:

#### Digital Libraries Initiative Homepage

<http://www.dli2.nsf.gov/>

*the definition of "digital library" continues to evolve...*

Internet accessible digital objects (representing text, data, documents, images, sound, video, agents, databases, middleware...) with sufficient identity, structure and contextual information to allow creating coherent collections on demand to service the needs of diverse user communities (query, analysis, communication, collaboration, ...)

*S. Griffin May 2000*