A Survey of Metadata Research for Organizing the Web

JANE L. HUNTER

ABSTRACT

THIS ARTICLE ATTEMPTS TO PROVIDE an overview of the key metadata research issues and the current projects and initiatives that are investigating methods and developing technologies aimed at improving our ability to discover, access, retrieve, and assimilate information on the Internet through the use of metadata.

1. INTRODUCTION

The rapid expansion of the Internet has led to a demand for systems and tools that can satisfy the more sophisticated requirements for storing, managing, searching, accessing, retrieving, sharing, and tracking complex resources of many different formats and media types.

Metadata is the value-added information that documents the administrative, descriptive, preservation, technical, and usage history and characteristics associated with resources. It provides the underlying foundation upon which digital asset management systems rely to provide fast, precise access to relevant resources across networks and between organizations. The metadata required to describe the highly heterogeneous, mixed-media objects on the Internet is infinitely more complex than simple metadata for resource discovery of textual documents through a library database. The problems and costs associated with generating and exploiting such metadata are correspondingly magnified.

Metadata standards, such as Dublin Core, provide a limited level of interoperability between systems and organizations to enable simple resource discovery. But, there are still many problems and issues that remain

Jane L. Hunter, Distributed Systems Technology Centre Pty. Ltd., Level 7, GP South, University of Queensland, St. Lucia, Queensland 4072, Australia
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 318–344

to be solved. Cory Doctorow (2001) believes that the vision of an Internet in which everyone describes their goods, services, or information using concise, accurate, and common or standardized metadata that is universally understood by both machines and humans is a "pipe-dream, founded on self-delusion, nerd hubris and hysterically inflated market opportunities." Other people cite the popularity and efficiency of Google as an example of an extremely successful search engine that does not depend on expensive and unreliable metadata. Google combines PageRanking (in which the relative importance of a document is measured by the number of links to it) with sophisticated text-matching techniques to retrieve precise, relevant, and comprehensive search results (Brin & Page, 1998).

Some of the major disadvantages of metadata are cost, unreliability, subjectivity, lack of authentication, and lack of interoperability with respect to syntax, semantics, vocabularies, languages, and underlying models. However, there are many researchers currently investigating strategies to overcome different aspects of these limitations in an effort to provide more efficient means of organizing content on the Internet. Other researchers are investigating metadata to describe the new types of real-time streaming content being generated by emerging broadband and wireless applications to enable both push and pull of this content based on users' needs. The goal of this article is to provide an overview of some of the key metadata research underway that is expected to improve our ability to search, discover, retrieve, and assimilate relevant information on the Internet regardless of the domain or format.

2. The Key Research Areas

In this section I have identified what I consider to be some of the key metadata research areas, both now and over the next few years. The following subsections provide a brief description of the work being undertaken and some key citations for each of the research areas summarized in the list below:

- Extensible Markup Language (XML)—XML and its associated technologies—XML Namespaces, XML Query languages, and XML Databases—are enabling implementers to develop metadata application profiles (XML Schemas) that combine metadata terms from different namespaces to satisfy the needs of a particular community or application. Large-scale XML descriptions of content are being stored in XML Databases and can be queried using XML Query Language. These are key technologies to enabling the automated computer processing, integration, and exchange of information.
- Semantic Web technologies—"The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee,

Hendler, & Lassila, 2001). There are two main building blocks for the semantic Web:

  - Formal languages—RDF (Resource Description Framework), DAML+OIL, and OWL (Web Ontology Language), which is being developed by the Web Ontology Working Group of the W3C.
  - Ontologies—communities will use the formal languages to define both domain-specific ontologies and top-level ontologies to enable relationships between ontologies to be determined for cross-domain searching, exchange, and information integration.
- Web Services—using open standards such as WSML, UDDI, and SOAP, Web services will enable the building of software applications without having to know who the users are, where they are, or anything else about them.
- Metadata Harvesting—the Open Archives Initiative (OAI) provides a protocol for data providers to make their metadata and content accessible—enabling value-added search and retrieval services to be built on top of harvested metadata.
- Multimedia metadata—there will be a further move away from textual resources to new multimedia formats that support better quality and higher compression ratios, e.g., images (JPEG-2000), video (MPEG-4), audio (MP3), 3D (VRML, Web3D), multimedia (SMIL, Shockwave Flash), and interactive digital objects. All of these new media types will require complex fine-grained metadata, extracted automatically where possible.
- Rights metadata—new emerging standards such as MPEG-21 and XrML are designed to enable automated copyright management and services.
- Automatic metadata extraction—technologies to enable the automatic classification and segmentation of digital resources. In particular, automatic image processing, speech recognition, and video-segmentation tools will enable content-based querying and retrieval of audiovisual content.
- Search engines:
  - Smarter agent-based search engines;
  - Federated search engines;
  - Peer-to-peer search engines;
  - Multimedia search engines;
  - Multilingual search engines;
  - New search interfaces—search interfaces that present results graphically;
  - Automatic/dynamic aggregation and generation of search results into hypermedia and multimedia presentations.
- Personalization/customization—autonomous agents that push relevant information to the user based on user preferences that may be personally configured or learned by the system.

- Broadband networks—multigigabit-capable networks for high-quality video-conferencing and visualization applications:
  - Grid computing—distributed computing and communications infrastructures for data intensive computing applications;
  - The Semantic Grid—the combination of semantic Web technologies with grid computing to provide large scale data access and integration to the e-Science community.
- Mobile and wireless technologies—delivery of information to mobile devices or appliances based on users' current context or location.
- Authentication—technologies to ensure trust and record the provenance of metadata.
- Annotation systems—enable users to attach their own subjective notes, opinions, and views to resources for others to access and read.
- Preservation metadata—metadata to support long-term preservation strategies for all types of digital resources.

2.1 XML Technologies and Metadata

XML and its associated technologies—XML Namespaces, XML Query languages, and XML Databases—are enabling implementers to develop metadata schemas, application profiles, large repositories of XML metadata, and search interfaces using XML Query Language. These technologies are key to enabling the automated computer-processing, integration, and exchange of information over the Internet.

2.1.1 Extensible Markup Language (XML). XML (W3C XML, 2003) is a simple, very flexible text format derived from SGML (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML is playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere. Because XML makes it possible to exchange data in a standard format, independent of storage, it has become the de facto standard for representing metadata descriptions of resources on the Internet.

2.1.2 XML Schema Language. XML Schema Language (W3C XML Schema, 2003) provides a means for defining the structure, content, and semantics of XML documents. It provides an inventory of XML markup constructs, which can constrain and document the meaning, usage, and relationships of the constituents of a class of XML documents: datatypes, elements and their content, attributes and their values, entities and their contents, and notations. Thus, the XML Schema Language can be used to define, describe, and catalog XML vocabularies for classes of XML documents, such as metadata descriptions of Web resources or digital objects.

XML Schemas have been used to define metadata schemas for a number of specific domains or applications—such as METS (Library of Congress, 2003), MPEG-7 (Martinez, 2002), MPEG-21 (Bormans & Hill, 2002), and NewsML (IPTC, 2001). An additional major metadata development

has been the employment of W3C's XML Schemas and XML Namespaces to combine metadata elements from different domains/namespaces into "application profiles" or metadata schemas that have been optimized for a particular application. For example, a particular community may want to combine elements of Dublin Core (DCMI, 2003), MPEG-7 (Martinez, 2002), and IMS (IMS, 2003) to enable the resource discovery of audio-visual learning objects.

2.1.3 XML Query. The mission of the XML Query Working Group (W3C XML Query, 2003) is to provide flexible query facilities to extract data from real and virtual documents on the Web, thereby providing the needed interaction between the Web world and the database world. Ultimately, collections of XML files will be accessed like databases. The new query language, XQuery, is still evolving, but it will provide a functional language comprised of several kinds of expressions that can be nested or composed with full generality. A working draft version of XQuery and a list of current XQuery implementations is available at http://www.w3.org/XML/Query.html.

2.1.4 XML Databases. There is a large amount of research and development going on in the area of XML databases. Ronald Bourret provides an excellent overview of the current state of this work and a comparison of current XML database technologies (Bourret, 2003a; Bourret, 2003b). Bourret divides XML Database solutions into the following categories:

- Middleware—software you call from your application to transfer data between XML documents and databases;
- XML-enabled databases—databases with extensions for transferring data between XML documents and themselves;
- Native XML databases—databases that store XML in "native" form, generally as some variant of the DOM mapped to an underlying data store. This includes the category formerly known as persistent DOM (PDOM) implementations;
- XML servers—XML-aware J2EE servers, Web application servers, integration engines, and custom servers. Some of these are used to build distributed applications while others are used simply to publish XML documents to the Web. Includes the category formerly known as XML application servers;
- Content Management Systems (CMS)—applications built on top of native XML databases and/or the file system for content/document management and which include features such as check-in/check-out, versioning, and editors;
- XML query engines—standalone engines that can query XML documents;
- XML data binding—products that can bind XML documents to objects. Some of these can also store/retrieve objects from the database.

2.1.5 Metadata Schema Registries. A number of groups have been tackling the issue of establishing registries of metadata schemas to enable the reuse and sharing of metadata vocabularies and to facilitate semantic interoperability. In particular the CORES project (CORES, 2003), which builds on the work of SCHEMAS (SCHEMAS, 2002), is exploring the use of metadata schema registries in order to enable the reuse of existing schemas, vocabularies, and application profiles that have been "registered."

2.2 The Semantic Web and Interoperability

According to Tim Berners-Lee, director of the World Wide Web Consortium (W3C), the Semantic Web is "an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.... The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users" (Berners-Lee, Hendler, & Lassila, 2001). But the Semantic Web has a long way to go before this dream is realized. The real power of the Semantic Web will be realized when programs and applications are created that collect Web content from diverse sources, process the information, and exchange the results with other programs.

Two of the key technological building blocks for the Semantic Web are:

- Formal languages for expressing semantics, such as the Resource Description Framework (RDF), DAML+OIL, and OWL (Web Ontology Language), which have been/are being developed within the W3C's Semantic Web Activity (W3C Semantic Web Activity, 2002); and
- The ontologies that are being constructed from such languages.

2.2.1 Formal Languages: RDF, DAML+OIL, OWL. The general consensus appears to be that while XML documents and schemas are ideal for defining the structural, formatting, and encoding constraints for a particular domain's metadata scheme, a different type of language is required for defining meaning or semantics.

The Resource Description Framework (RDF) (W3C, RDF Syntax, & Model Recommendation, 1999; W3C RDF Vocabulary Description Language, 2003) uses triples to makes assertions that particular things (people, Web pages, or whatever) have properties (such as "is a sister of," "is the author of") with certain values (another person, another Web page). The triples of RDF form webs of information about related things. Because RDF uses URIs to encode this information in a document, the URIs ensure that concepts are not just words in a document but are tied to a unique definition that everyone can find on the Web. This work is being undertaken by the RDF Core Working Group of the W3C.

The W3C Web Ontology Working Group (W3C Web Ontology, 2003) is building upon the RDF Core work to develop a language for defining structured Web-based ontologies that will provide richer integration and interoperability of data among descriptive communities. This is the Web Ontology Language (OWL) (W3C, OWL, 2003), which in turn is building upon the DAML+OIL (DAML+OIL, 2001) specification developed by DARPA.

2.2.2 Ontologies. An ontology consists of a set of concepts, axioms, and relationships that describes a domain of interest. An ontology is similar to a dictionary or glossary but with greater detail and structure and expressed in a formal language (e.g., OWL) that enables computers to process its content. Ontologies can enhance the functioning of the Web to improve the accuracy of Web searches and to relate the information in a resource to the associated knowledge structures and inference rules defined in the ontology.

Upper ontologies provide a structure and a set of general concepts upon which domain-specific ontologies (e.g., medical, financial, engineering, sports, etc.) could be constructed. An upper ontology is limited to concepts that are abstract and generic enough to address a broad range of domain areas at a high level. Computers utilize upper ontologies for applications such as data interoperability, information search and retrieval, automated inferencing, and natural language processing.

A number of research and standards groups are working on the development of common conceptual models (or upper ontologies) to facilitate interoperability between metadata vocabularies and the integration of information from different domains. The Harmony project developed the ABC Ontology/Model (Lagoze & Hunter, 2001)—a top-level ontology to facilitate interoperability between metadata schemas within the digital library domain. The CIDOC CRM (CIDOC CRM, 2003) has been developed to facilitate information exchange in the cultural heritage and museum community. The Standard Upper Ontology (SUO, 2002) is being developed by the IEEE SUO Working Group.

Many communities are developing domain-specific or application-specific ontologies. Some examples include biomedical ontologies such as OpenGALEN (OpenGALEN, 2002) and SNOMED CT (SNOMED CT, 2003), financial, and sporting ontologies such as the soccer, baseball, or running ontologies in the DAML Ontology Library (DAML Ontology Library, 2003).

A large number of research efforts are focusing on the development of tools for building and editing ontologies (Denny, 2002)—these are moving towards collaborative tools such as OntoEdit (Sure et al., 2002) and built-in support for RuleML to enable the specification of inferencing rules.

2.2.4 Topic Maps. Topic Maps (Topic Maps, 2000) is a new ISO standard for a system describing knowledge structures and associating them with

information resources. They provide powerful ways of navigating large and interconnected corpora. Instead of replicating the features of a book index, the topic map generalizes them, extending them in many directions at once. The difference between Topic Maps and RDF is that Topic Maps are centered on topics while RDF is centered on resources. RDF annotates the resources directly whilst topic maps create a "virtual map" above the resources, leaving them unchanged.

2.2.5 Ontology Storage and Querying. A number of research groups are currently working on the development of inferencing tools and deductive query engines to enable the deduction of new information or knowledge from assertions or metadata and ontologies expressed in formal ontology languages (RDF, DAML+OIL, or OWL). A technical report on "Ontology Storage and Querying," published recently by ICS FORTH in Crete, provides a very good survey of the current state of ontology storage and querying tools (Magkanaraki et al., 2002).

2.3 Web Services

Web services (W3C Web Services Activity, 2003) are a relatively new concept, expected to evolve rapidly over the next few years. They could be the first major practical manifestation of Semantic Web–based thinking. Detailed definitions vary, but Web services will enable the building of software applications without having to know who the users are, where they are, or anything else about them. In the next few years, Web services may be developed that can be understood and used automatically by the computing devices of users and of public libraries. External Application Services Providers (ASPs) may also provide such services. Web services are based on open, Internet standards. The core standards and protocols for Web services are being developed and are expected to be finalized by 2003. They include (in addition to XML):

- Web Services Description Language (WSDL) (WSDL, 2003), which enables a common description of Web Services;
- Universal Description, Discovery, and Integration (UDDI) (OASIS, 2003) registries, which expose information about a business or other entity and its technical interfaces;
- Simple Object Access Protocol (SOAP)/XML Protocol (W3C XML Protocol Working Group, 2003), which enables structured message exchange between computer programs.

The concept of Web services is currently being developed under the banner of e-commerce. However, there do appear to be potential applications for public sector service providers. For example, search interfaces could be accessed or provided as Web services by public libraries or by Application Service Providers (ASPs) on their behalf.

2.4 Metadata Harvesting—The Open Archives Initiative (OAI)

The Open Archives Initiative (OAI) (OAI, 2003) is a community that has defined an interoperability framework, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), to facilitate the sharing of metadata. Using this protocol, data providers are able to make metadata about their collections available for harvesting through an HTTP-based protocol. Service providers then use this metadata to create value-added services. OAI-PMH Version 2.0 was released in February 2003 (OAI-PMH, 2003).

To facilitate interoperability, data providers are required to supply metadata that complies to a common schema, the unqualified Dublin Core Metadata Element Set. Additional schemas are also allowed and are distinguished through the use of a metadata prefix.

Although originating in the E-Print community, OAI data providers now include a number of multimedia collections such as the Library of Congress American Memory collection (Library of Congress, 2002), OpenVideo (OpenVideo, 2002), and University of Illinois historical images (UIL, 2002). DSpace at MIT (DSpace, 2002) is also a registered data provider. HP Labs and MIT Libraries have also made the DSpace software available—it is an open-source, digital asset management software platform that enables institutions to capture and describe digital works using a submission workflow module; distribute an institution's digital works over the Web through a search and retrieval system; and store and preserve digital works over the long term. And it supports OAI-PMH Version 2.0.

To date, OAI service providers have mostly developed simple search and retrieval services (OAI Registered Service Providers, 2002). These include Arc, citebaseSearch, and my.OAI. Scirius searches and retrieves specifically scientific data—from the Web, proprietary databases, and Open Archives. One of the more interesting services is DP9, a gateway service that allows traditional Web search engines (e.g., Google) to index otherwise hidden information from OAI archives. The DSTC's MAENAD project developed a search, retrieval, and presentation system for OAI that searches for and retrieves mixed-media resources on a particular topic, determines the semantic relationships between the retrieved objects, and combines them into a coherent multimedia presentation, based on their relationships to each other (Little, Guerts, & Hunter, 2002).

2.5 Multimedia Metadata

Audiovisual resources in the form of still pictures, graphics, 3D models, audio, speech, and video will play an increasingly pervasive role in our lives and, because of the complex information-rich nature of such content, value-added services such as analysis, interpretation, and metadata creation become much more difficult, subjective, time consuming, and expensive. Audiovisual content requires some level of computational interpretation

and processing in order to generate metadata of useful granularity efficiently. Standardized multimedia metadata representations that will allow some degree of machine interpretation will be necessary. The MPEG-7 and MPEG-21 standards have been developed to support such requirements.

2.5.1 MPEG-7 Multimedia Content Description Interface. MPEG-7 (Martinez, 2002), the "Multimedia Content Description Interface," is an ISO/IEC standard for describing multimedia content, developed by the Moving Pictures Expert Group (MPEG). The goal of this standard is to provide a rich set of standardized tools to enable both humans and machines to generate and understand audiovisual descriptions that can be used to enable fast, efficient retrieval from digital archives (pull applications) as well as filtering of streamed audiovisual broadcasts on the Internet (push applications). MPEG-7 can describe audiovisual information regardless of storage, coding, display, transmission, medium, or technology. It addresses a wide variety of media types including still pictures, graphics, 3D models, audio, speech, video, and combinations of these (e.g., multimedia presentations). The MPEG-7 specification provides:

- A core set of Descriptors (Ds) that can be used to describe the various features of multimedia content;
- Predefined structures of Descriptors and their relationships, called Description Schemes (DSs).

MPEG-7 Multimedia Description Schemes enable descriptions of multimedia content, including:

- Information describing the creation and production processes of the content (director, title, short feature movie);
- Information related to the usage of the content (copyright pointers, usage history, broadcast schedule);
- Media information on the storage features of the content (storage format, encoding);
- Structural information on spatial, temporal, or spatio-temporal components of the content (scene cuts, segmentation in regions, region motion tracking);
- Information about low-level features in the content (colors, textures, sound timbres, melody description);
- Conceptual, semantic information of the reality captured by the content (objects and events, interactions among objects);
- Information about how to browse the content in an efficient way (summaries, views, variations, spatial and frequency sub-bands);
- Organization information about collections of objects and models that allow multimedia content to be characterized on the basis of probabilities, statistics, and examples;

- Information about the interaction of the user with the content (user preferences, usage history).

Until now research in this area has primarily focused on developing efficient, low-level, digital signal processing methods to extract values for image, video, and audio Descriptors such as color, shape, texture, motion, volume, and phonemes. Algorithms have been developed to automatically segment video into scenes and shots for faster browsing and retrieval or to automatically transcribe speech and video content. Multimedia metadata research is now focusing on how to automatically generate semantic descriptions of multimedia (machine recognition of objects and events) from combinations of low-level descriptors such as color, texture, and shape and audio descriptors to enable natural language querying and higher-level knowledge extraction.

Additional research efforts are investigating how to combine ontologies for specific domains, e.g., sports, medical, bio-informatics, and nanotechnology with MPEG-7 to describe multimedia content in terms relevant to the particular domain or to relate and integrate multimedia information from across domains or disciplines.

2.5.2 MPEG-21—Multimedia Framework. The goal of MPEG's latest initiative, MPEG-21 (ISO/IEC 18034–1) (Bormans & Hill, 2002), the Multimedia Framework, is to define the technology needed to support *Users* to exchange, access, consume, trade, and otherwise manipulate multimedia *Digital Items* in an efficient, transparent, and interoperable way. *Users* may be content creators, producers, distributors, service providers, or consumers. They include individuals, communities, organizations, corporations, consortia, governments, and other standards bodies and initiatives around the world. The fundamental unit of content is called the *Digital Item,* and it could be anything from a textual document or a simple Web page to a video collection or a music album.

At its most basic level, MPEG-21 provides a framework in which one *User* interacts with another *User* and the object of that interaction is a *Digital Item* commonly called content. Some such interactions are creating content, providing content, archiving content, rating content, enhancing and delivering content, aggregating content, delivering content, syndicating content, retail selling of content, consuming content, subscribing to content, regulating content, facilitating transactions that occur from any of the above, and regulating transactions that occur from any of the above.

The current MPEG-21 Work Plan consists of nine parts:

- Part 1: Vision, Technologies, and Strategies—a technical report that describes MPEG-21's architectural elements together with the functional requirements for their specification;

- Part 2—Digital Item Declaration—a flexible model for precisely defining the scope and components of a Digital Item;
- Part 3—Digital Item Identification—a specification for uniquely identifying Digital Items and their components;
- Part 4—Intellectual Property Management and Protection (IPMP)—to provide interoperability between IPMP tools, such as MPEG-4's IPMP hooks;
- Part 5—Rights Expression Language—a machine-readable language that can declare rights and permissions using the terms as defined in the Rights Data Dictionary (XrML);
- Part 6—Rights Data Dictionary—definitions of terms to support Part 5;
- Part 7—Digital Item Adaptation—adaptation may be based on user, terminal, network and environmental characteristics, resource adaptability, or session mobility;
- Part 8—Reference Software—used to test conformance with requirements and the standard's specifications;
- Part 9—File Format—this is expected to inherit many MPEG-4 concepts, since it will need to be able to encapsulate digital item information, still and dynamic media, metadata, and layout data in both textual and binary forms.

Future work plans for MPEG-21 include developing functional requirements and solutions to the persistent association of identification and description with Digital Items; scalable, error-resilient content representation; and the accurate recording of all events.

2.6 Rights Metadata

The Internet has been characterized as the largest threat to copyright since its inception. Copyrighted works on the Internet include news stories, software, novels, screenplays, graphics, pictures, usenet messages, and even e-mail. The reality is that almost everything on the Internet is protected by copyright law. This can pose problems for both hapless surfers as well as the copyright owners.

A number of XML-based vocabularies have been developed to define the usage and access rights associated with digital resources—XrML (XrML, 2003), developed by ContentGuard, and ODRL (ODRL, 2003), developed by IPR Systems are the two major contenders. XrML has been adopted by MPEG-21 as its Rights Expression Language, and ODRL was recently selected by the Open Mobile Alliance as its rights language for mobile content.

In addition there are a number of researchers investigating the development of well-defined, underlying, interoperable data models for rights management that is necessary for facilitating interoperability and the integration of information (indecs Framework, 2000; Delgado et al., 2002).

Project RoMEO (Rights MEtadata for Open archiving) (RoMEO, 2003) is investigating the rights issues surrounding the "self-archiving" of research in the U.K. academic community under the Open Archive Initiative's Protocol for Metadata Harvesting. Academic and self-publishing authors who make their works available through Open Archives are more concerned with issues such as plagiarism, corruption, or misuse of the text than financial returns to the author or publisher.

The "Indigenous Collections Management Project" being undertaken by Distributed Systems Technology Centre (DSTC), University of Queensland, in collaboration with the Smithsonian's National Museum of the American Indian, has also been investigating metadata for the rights management and protection of traditional knowledge belonging to indigenous communities, in accordance with customary laws regarding access (Hunter, 2002; Hunter, Koopman, & Sledge, 2003).

2.7 Automatic Metadata Extraction

Because of the high cost and subjectivity associated with human-generated metadata, a large number of research initiatives are focusing on technologies to enable the automatic classification and segmentation of digital resources—i.e., computer-generated metadata for textual documents, images, audio, and video resources.

2.7.1 Automatic Document Indexing/Classification. Automatic-categorization software (Reamy, 2002) uses a wide variety of techniques to assign documents into subject categories. Techniques include statistical Bayesian analysis of the patterns of words in the document; clustering of sets of documents based on similarities; advanced vector machines that represent every word and its frequency with a vector; neural networks; sophisticated linguistic inferences; the use of preexisting sets of categories; and seeding categories with keywords. The most common method used by autocategorization software is to scan every word in a document and analyze the frequencies of patterns of words and, based on a comparison with an existing taxonomy, assign the document to a particular category in the taxonomy. Other approaches use "clustering" or "taxonomy building" in which the software is pointed at a collection of documents (e.g., 10,000–100,000) and it searches through all the combinations of words to find clumps or clusters of documents that appear to belong together. Some systems are capable of automatically generating a summary of a document by scanning through the document and finding important sentences using rules like the first sentence of the first paragraph is often important. Another common feature of autocategorization is noun phrase extraction—the extracted list of noun phrases can be used to generate a catalog of entities covered by the collection.

Autocategorization cannot completely replace a librarian or information architect, although it can make them more productive, save them

time, and produce a better end-product. The software itself, without some human rules–based categorization, cannot currently achieve more than about 90 percent accuracy. While it is much faster than a human categorizer, it is still not as good as a human.

2.7.2 Image Indexing. Image retrieval research has moved on from the IBM QBIC (query by image content) system (QBIC, 2001), which uses colors, textures, and shapes to search for images. New research is focusing on semantics-sensitive matching (DCSE, 2003; Barnard, 2003) and automatic linguistic indexing (Wang & Li, 2003), in which the system is capable of recognizing real-world objects or concepts.

2.7.3 Speech Indexing and Retrieval. Speech recognition is increasingly being applied to the indexing and retrieval of digitized speech archives. Dragon Systems (Dragon Systems, 2003) has developed a system that creates a keyword index of spoken words from within volumes of recorded audio, eliminating the need to listen for hours to pinpoint information. Speech recognition systems can generate searchable text that is indexed to time code on the recorded media, so users can both call up text and jump right to the audio clip containing the keyword. Normally, running a speech recognizer on audio recordings doesn't produce a highly accurate transcript because speech-recognition systems have difficulty if they haven't been trained for a particular speaker or if the speech is continuous. However, the latest speech recognition systems will work even in noisy environments, are speaker-independent, work on continuous speech, and are able to separate two speakers talking at once. Dragon is also working on its own database for storing and retrieving audio indexes.

2.7.4 Natural Language and Spoken Language Querying. Dragon has also developed systems that allow users to retrieve information from databases using natural language queries. Such systems are expected to become more commonplace in the future (Oard, 2003).

2.7.5 Video Indexing and Retrieval. Commercial systems such as Virage (Virage, 2003), Convera (Convera Screening Room, 2003), and Artesia (Artesia, 2003) are capable of parsing hours of video, segmenting it, and turning it into an easily searchable and browsable database.

The latest video-indexing systems combine a number of indexing methods—embedded textual data, (SMPTE timecode, lineup files, and closed captions), scene change detection, visual clues, and continuous-speech recognition to convert spoken words into text. For example, CMU's Informedia project (Informedia, 2003) combines text, speech, image, and video recognition techniques to segment and index video archives and enable intelligent search and retrieval. The system can automatically analyze videos and extract named entities from transcripts, which can be used to produce time and location metadata. This metadata can then be used to explore archives dynamically using temporal and spatial graphical user interfaces, e.g., mapping interfaces or date sliders. For example—"give me

all video content on air crashes in South America in early 2000" (Ng et al., 2003).

Current research in this field is concentrating on the difficult problem of extracting metadata in real-time from streaming video content, rather than during a postprocessing step.

2.8 Search Engine Research and Development

2.8.1 Smarter Agent-based Search Engines. One of the major advances in search engines in the future will be in the use of "intelligent agents" and expert systems that apply artificial intelligence (AI), ontologies, and knowledge bases to enable all relevant information on a particular subject to be retrieved and integrated. Improved user interfaces will become available through the incorporation of expert systems into online catalog searching, i.e., "intelligent" sophisticated online systems that incorporate AI, knowledge bases, and ontologies. In the future librarians will use "intelligent agent kits" that will crawl over the Web retrieving relevant information and will analyze and interpret it to create a body of knowledge for a specific purpose. Periodic resampling will automatically keep it up-to-date. However, human intervention will still be needed to customize, supervise, and check the computer-generated results (Virginia Tech, 1997; Nardi & O'Day, 1998).

2.8.2 Federated Search Engines. Quite a large number of metadata research projects are focusing on the problems of federated searching across distributed, heterogeneous, networked digital libraries and the interoperability problems that need to be overcome (Gonçalves et al., 2001; Liu et al., 2002). For example, the MetaLib project, at the University of East Anglia, implements a single integrated environment and cross-searching portal for managing and searching electronic resources, whether these be abstracting and indexing databases, full-text e-journal services, CD-ROMs, library catalogs, information gateways, or local collections (Lewis, 2002).

2.8.3 Peer-to-Peer JXTA-based Search Engines. Peer-to-peer (P2P) search engines are based on the idea of decentralized metadata provided by networked peers rather than clients accessing centralized metadata repositories sitting on a server. Sam Joseph at the University of Tokyo has written an excellent overview of Internet search engines based on decentralized metadata (Joseph, 2003).

JXTA (short for Jxtapose) is a peer-to-peer interoperability framework created by Sun. It incorporates a number of protocols, but the most relevant to the idea of decentralized metadata is the Peer Discovery Protocol (PDP). PDP allows a peer to advertise its own resources and discover the resources from other peers. Every peer resource is described and published using an advertisement, which is an XML document that describes a network resource. JXTASearch operates over the lower-level JXTA protocols (JXTA, 2003).

Edutella (Edutella, 2002) is an RDF-based Metadata Infrastructure for P2P Applications based on JXTA. The first application developed by Edutella focuses a P2P network for the exchange of educational resources between German universities (including Hannover, Braunschweig, and Karlsruhe), Swedish universities (including Stockholm and Uppsala), Stanford University, and others.

2.8.4 Multimedia Search Engines. More and more search engines are becoming multimedia-capable—even allowing users to specify media types (images, video, or audio) and formats (e.g., JPEG, MP3, SMIL). Examples include the FAST Multimedia Search Engine (FAST, 2000), Alta Vista (AltaVista, 2003), Google Image Search (Google, 2003), Singingfish Multimedia Search (SingingFish, 2002), Friskit Music Streaming Media Search (Friskit, 2002), and the Fossick Online Multimedia and Digital Image Search (Fossick, 2003).

2.8.5 Cross-lingual Search Engines. In the future, universal translators will automatically translate a query in one particular language into any number of other languages and also translate the results into the original query language. There are a number of research projects and search engines focusing on cross-lingual search engines, e.g., SPIRIT-W3, a distributed cross-lingual indexing and search engine (Fluhr et al., 1997), and the TITAN Cross-Language Web search engine (TITAN, 2003).

2.9 Graphical/Multimedia Presentation of Results

2.9.1 Graphical Presentation of Search Results. More search engines are going to present search results in more innovative graphical ways other than simple lists of URLs. Interfaces like Kartoo (Kartoo, 2000) and WebBrain (WebBrain, 2001) illustrate the relationships between retrieved digital resources graphically. Kartoo uses Flash to provide a graphical representation of the results. The results are displayed in a 2–3D map representing sites that match your query as nodes on the map, and relationships between nodes are represented as labeled arcs. WebBrain presents search results in a graphical browse interface that allows users to navigate through related topics.

TouchGraph GoogleBrowser (TouchGraph, 2001) is a tool for visually browsing the Google database by exploring links between related sites. It uses Google's database to determine and display the linkages between a URL that you enter and other pages on the Web. Results are displayed as a graph, showing both inbound and outbound relationships between URLs.

"Friend of a Friend" or *foaf* (foaf, 2000) is an RDF vocabulary for describing the relationships between people, invented by Dan Brickley and Libby Miller of RDF Web. foafCORP (foafCORP, 2002) is an interesting semantic Web visualization of the interconnectedness of corporate America based on the foaf RDF vocabulary. It provides a simple graphical user

interface to trace relationships between board members of major companies in the United States.

2.9.2 Automatic Aggregation/Compilation Tools. The rapid growth in multimedia content on the Internet, the standardization of machine-processable, semantically rich (RDF-based) content descriptions, and the ability to perform semantic inferencing have together led to the development of systems that can automatically retrieve and aggregate semantically related multimedia objects and generate intelligent multimedia presentations on a particular topic, i.e., knowledge-based authoring tools (Little et al., 2002; CWI, 2000; Conlan et al., 2000; André, 2000).

Automatic information aggregation tools that can dynamically generate hypermedia and multimedia learning objects will be extremely relevant to libraries in the future. Such tools will expedite the cost-effective creation of value-added learning objects and will also ensure that any relevant content only recently made available by content providers will be automatically incorporated in the dynamically generated learning objects.

2.10 Metadata for Personalization/Customization

The individualization of information, based on users' needs, abilities, prior learning, interests, context, etc., is a major metadata-related research issue (Lynch, 2001a). The ability to push relevant, dynamically generated information to the user, based on user preferences, may be implemented

- either by explicit user input of their preferences;
- or learned by the system by tracking usage patterns and preferences and adapting the system and interfaces accordingly.

The idea is that users can get what they want without having to ask. The technologies involved in recommender systems are information filtering, collaborated filtering, user profiling, machine learning, case-based retrieval, data mining, and similarity-based retrieval. User preferences typically include information such as the user's name, age, prior learning, learning style, topics of interest, language, subscriptions, device capabilities, media choice, rights broker, payment information, etc. Manually entering this information will produce better results than system-generated preferences, but it is time consuming and expensive. More advanced systems in the future will use automatic machine-learning techniques to determine users' interests and preferences dynamically rather than depending on user input.

Some examples of "personalized current awareness news services" are Net2one (Net2one, 2003), MSNBC News Filters (MSNBC, 2003), and the eLib Newsagent project (eLib Newsagent, 2000). These services allow users to define their interests and then receive daily updated relevant reports. Filtering of Web radio and TV broadcasts will also be possible in the future, based on users' specifications of their interests and the embedding of stan-

dardized content descriptions, such as MPEG-7, within the video streams (Rogers et al., 2002).

2.11 Metadata for Broadband/Grid Applications

The delivery and integration of information is shifting to wireless mobile devices and high-performance broadband networks. To support research and development in advanced grid and networking services and applications, a number of broadband multigigabit advanced networks have been established throughout the world and made accessible to the research and higher education communities of these regions:

- Internet2—U.S. broadband research network (Internet2, 2003);
- GrangeNet—Australian broadband network (GrangeNet, 2003);
- Canarie—Canadian broadband network (Canarie, 2002);
- DANTE—European broadband research network (DANTE, 2003);
- APAN—Asia Pacific Advanced Network (APAN, 2003).

Related research projects are focusing on real-time, collaborative, distributed applications that require very high-quality video or high-speed access to large data sets for remote collaboration and visualization. Examples of applications include remote telemicroscopy, remote surgery, 3D visualization of large datasets (e.g., bio-informatics, astronomy data), collaborative editing of HDTV-quality digital video, and distributed real-time music and dance performances.

2.11.1 Grid Computing. Computational Grids enable the sharing, selection, and aggregation of a wide variety of geographically distributed computational resources (such as supercomputers, computer clusters, storage systems, data sources, instruments, people) and presents them as a single, unified resource for solving large-scale compute and data-intensive computing applications (e.g., molecular modeling for drug design, brain activity analysis, climate modeling, and high-energy physics) (Grid Computing, 2000). Wide-area distributed computing, or "grid" technologies, provide the foundation to a number of large-scale efforts utilizing the global Internet to build distributed computing and communications infrastructures. A list of current grid initiatives and projects can be found at http://www.gridforum.org/L_Involved_Mktg/init.htm (GGF, 2003).

2.11.2 The Semantic Grid. This term refers to the underlying computer infrastructure needed to support scientists who want to generate, analyze, share, and discuss their results/data over broadband Grid networks—basically it is the combination of Semantic Web technologies with Grid computing for the scientific community (Semantic Grid, 2003).

In particular, the combination of Semantic Web technologies with live information flows is highly relevant to grid computing and is an emerging research area—for example, the multiplexing (embedding) of live metadata

with multicast video streams raises the issue of Quality of Service (QoS) demands on the network.

Archival and indexing tools for collaborative video conferences held through Access Grid Nodes are going to be in demand. In typical access grid installations, there are three displays with multiple views. There is a live exchange of information. Events such as remote camera control and slide transitions could be used to segment and index the meetings for later search and browsing. Notes and annotations taken during the meeting provide additional sets of metadata that can be stored and shared. Metadata schemes to support collaborative meetings and collaboratories will be required.

Scientists collaborating on grid networks are going to require methods and tools to build large-scale ontologies, annotation services, inference engines, integration tools, and knowledge discovery services for Grid and e-Science applications (De Roure et al., 2001).

2.12 Metadata for Wireless Applications

Infrared detection and transmission can be used in libraries to beam context-sensitive data or applications to users' PDAs, depending on where they are physically located (Kaine-Krolak & Novak, 1995). Similarly, GPS information can be used to download location-relevant data to users' PDAs or laptops when they are traveling, e.g., scientists on field trips. Such context-sensitive applications require location metadata to be attached to information resources in databases connected to wireless networks.

The ROADNet (ROADNet, 2002) project on HPWREN (HPWREN, 2001), a high-performance wireless network, is a demonstration of the collection and streaming of real-time seismic, oceanographic, hydrological, ecological, geodetic, and physical data and metadata via a wireless network. Real-time numeric, audio, and video data are collected via field sensors and researchers connected to HPWREN and posted to discipline-specific servers connected over a network. This data is immediately accessible by interdisciplinary scientists in near–real time. Extraction of metadata from real-time data flow, as well as high-speed metadata fusion across multiple data sensors, are high-priority research goals within applications such as ROADNet.

2.13 Metadata Authentication

Manually generated metadata for Web resources cannot be assumed to be accurate or precise descriptions of those resources. The metadata and/or the Web page may have been deliberately constructed or edited so as to misrepresent the content of the resource and to manipulate the behavior of the retrieval systems that use the metadata. Basically, anyone can create any metadata they want about any object on the Internet with

any motivation. There is an urgent need for technologies that can vouch for or authenticate metadata so that Web indexing systems that crawl across the Internet developing Web index databases know when the associated metadata can be trusted (Lynch, 2001b).

Hence there are a number of research projects investigating methods for explicitly identifying and validating the source of metadata assertions, using technologies such as XML Signature. Search engines give higher confidence weightings to metadata signed by trusted providers, and this is reflected in the retrieved search results.

The XML Signature Working Group, a joint working group of the IETF and W3C (W3C XML Signature, 2003), has developed an XML compliant syntax for representing signatures of Web resources (or anything referenceable by a URI) and procedures for computing and verifying such signatures. Such signatures can easily be applied to metadata and used by Web servers and search engines to ensure metadata's authenticity and integrity. The XML Signature specification is based on Public Key Cryptography in which signed and protected data is transformed according to an algorithm parameterized by a pair of numbers—the so-called public and private keys. Public Key Infrastructure (PKI) systems provide management services for key registries—they bind users' identities to digital certificates and public/private key pairs that have been assigned and warranted by trusted third parties (Certificate Authorities).

Another approach is the Pretty Good Privacy (PGP) system (PGP, 2002) in which a "Web of Trust" is built up from an established list of known and trusted identity/key bindings. Trust is established in new unfamiliar identity/key bindings because they are cryptographically signed by one or more parties that are already trusted.

2.14 Annotation Systems

The motivation behind annotation systems is related to the issue of metadata trust and authentication—users can attach their own metadata, views, opinions, comments, ratings, and recommendations to particular resources or documents on the Web, which can be read and shared with others. The basic philosophy is that we are more likely to value and trust the opinions of people we respect than metadata of unknown origin.

The W3C's Annotea system (W3C Annotea, 2001) and DARPA's Web Annotation Service (DARPA, 1998) are two Web-based annotation systems that have been developed. Current research is focusing on annotation systems within real-time collaborative environments (Benz and Lijding, 1998), annotation tools for film/video and multimedia content (IBM VideoAnnEx, 2001; Ricoh MovieTool, 2002; ZGDV VIDETO, 2002; DSTC FilmEd, 2003), and tools to enable the attachment of spoken annotations to digital resources (PAXit, 2003) such as images or photographs.

2.15 Weblogging Metadata

Weblogging or Blogging (Sullivan, 2002; Reynolds et al., 2002) is a very successful paradigm for lightweight publishing, which has grown sharply in popularity over the past few years and is being used increasingly to facilitate communication and discussion within online communities. The idea of semantic blogging is to add additional semantic structure to items shared over blog channels or RSS feeds to enable semantic search, navigation, and filtering of blogs or streaming data.

Blizg (Blizg, 2003) and BlogChalking (BlogChalking, 2002) are two examples of Weblog search engines that use metadata to enable searching across Weblog archives and the detection of useful connections between and among blogs.

2.16 Metadata for Preservation

A number of initiatives have been focusing on the use of metadata to support the digital preservation of resources. Such initiatives include: Reference Model for an Open Archival Information System (OAIS, 2002), the CURL Exemplars in Digital Archives project (CEDARS, 2002), the National Library of Australia (NLA) PANDORA project (PANDORA, 2002), the Networked European Deposit Library (NEDLIB, 2001), and the Online Computer Library Center/Research Libraries Group (OCLC/RLG) Working Group on Preservation Metadata (OCLC/RLG, 2003).

These initiatives rely on the preservation of both the original bytestream/digital object, as well as detailed metadata that will enable the preserved data to be interpreted in the future. The preservation metadata provides sufficient technical information about the resources to support either migration or emulation. Metadata can facilitate the long-term access of the digital resources by providing a complete description of the technical environment needed to view the work, the applications and version numbers needed, and decompression schemes, as well as any other files that need to be linked to it. However, associating appropriate metadata with digital objects will require new workflows and metadata input tools at the points of creation, acquisition, reuse, migration, etc. This will demand initial effort to be made the first time a particular class of digital resource is received into a collection. However, assuming many of the same class of resource are received, economies of scale can be achieved by reusing the same metadata model and input tools.

The Library of Congress's Metadata Encoding and Transmission Standard (METS) (Library of Congress, 2003) schema provides a flexible mechanism for encoding descriptive, administrative, and structural metadata for a digital library object and for expressing the complex links between these various forms of metadata.

Other research initiatives are investigating extensions to METS to enable the preservation of audiovisual content or complex multimedia

objects such as multimedia artworks (Avant Garde, 2003; DSTC NewMedia, 2003). These approaches involve the association of ancillary and contextual information such as interviews with artists and the use of the Bit Stream Description Language (BSDL) (Amielh and Devillers, 2002) to convert objects preserved as bit streams into formats that can be displayed on the current platforms.

3. Conclusions

In this paper, I have attempted to provide an overview of some of the key metadata research efforts currently underway that are expected to improve our ability to search, discover, retrieve, and assimilate information on the Internet. The number and extent of the research projects and initiatives described in this paper demonstrate three things:

1. The resource requirements and intellectual and technical issues associated with metadata development, management, and exploitation are far from trivial, and we are still a long way from *MetaUtopia;*
2. Metadata means many different things to many different people, and its effectiveness depends on implementers resolving key issues, including:
   - Identifying the best metadata models, schemas, and vocabularies to satisfy their requirements;
   - Deciding on the granularity of metadata necessary for their needs—this will involve a trade-off between the costs of developing and managing metadata, the desired search capabilities, potential future uses, and preservation needs;
   - Balancing the costs and subjectivity of user-generated metadata with the anticipated error rate of automatic metadata extraction tools;
   - Ensuring the currency, authenticity, and integrity of the metadata;
   - Choosing between decentralized, distributed metadata architectures and centralized repositories for the storage and management of metadata.
3. Despite its problems, metadata is still considered a very useful and valuable component in organizing content on the Internet and in enabling us to find relevant information and services effectively.

References

Alta Vista. (2003). Retrieved July 28, 2003, from http://www.altavista.com/.

Amielh, M., & Devillers, S. (2002). Bitstream syntax description language: Application of XML-schema to multimedia content adaptation. WWW2002 Conference. Honolulu. Retrieved August 11, 2003, from http://www2002.org/CDROM/alternate/334/.

André, E. (2000). The generation of multimedia documents. In R. Dale, H. Moisl, and H. Somers (Eds.), *A handbook of natural language processing: Techniques and applications for the processing of language as text.* (pp. 305–327). Tampa: Marcel Dekker, Inc. Retrieved August 11, 2003, from http://www.dfki.de/imedia/papers/handbook.ps.

Artesia. (2003). Retrieved August 11, 2003, from http://www.artesiatech.com/.

Asia Pacific Advanced Network (APAN). (2003). Retrieved August 11, 2003, from http://apan.net/.

Avant Garde. (2003). Archiving the avant garde: Documenting and preserving variable MediaArt. Retrieved August 11, 2003, from http://www.bampfa.berkeley.edu/ciao/avant_garde.html.

Barnard, K. (2003). Computer vision meets digital libraries. Retrieved August 11, 2003, from http://elib.cs.berkeley.edu/vision.html.

Benz, H., & Lijding, M. E. (1998). Asynchronously replicated shared workspaces for a multimedia annotation service over Internet. *Lecture notes in computer science.* Retrieved August 11, 2003, from http://elib.uni-stuttgart.de/opus/volltexte/1999/533/.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The semantic Web. *Scientific American.* Retrieved August 11, 2003, from http://www.sciam.com/article.cfm?colID=1&articleID=00048144–10D2–1C70–84A9809EC588EF21.

Blizg. (2003). Retrieved August 11, 2003, from http://blizg.com/.

BlogChalking. (2002). Retrieved August 11, 2003, from http://www.blogchalking.tk/.

Bormans, J. & Hill, K. (2002). MPEG-21 overview V.5. Retrieved August 11, 2003, from http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm.

Bourret, R. (2003a). XML and databases. Retrieved August 11, 2003, from http://www.rpbourret.com/xml/XMLAndDatabases.htm.

Bourret, R. (2003b). XML Database products. Retrieved August 11, 2003, from http://www.rphourret.com/xml/XMLDatabaseProds.htm.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th International World Wide Web Conference (WWW7)* (pp 107–117). Brisbane, Australia. Retrieved August 11, 2003, from http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm.

Canarie. (2002). Retrieved August 11, 2003, from http://www.canarie.ca/.

CEDARS, CURL. (2002). Exemplars in digital archives. Retrieved August 11, 2003, from http://www.leeds.ac.uk/cedars/.

CIDOC CRM. (2003). CIDOC conceptual reference model. Retrieved August 11, 2003, from http://cidoc.ics.forth.gr/.

Conlan, O., Wade, V., Bruen, C., & Gargan, M. (2002). Multi-model, metadata driven approach to adaptive hypermedia, services for personalized e-learning. Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Malaga, Spain, May 2002.

Convera Screening Room. (2003). Retrieved August 11, 2003, from http://www.convera.com/Products/products_sr.asp.

CORES. (2003). CORES—A forum on shared metadata vocabularies. Retrieved August 11, 2003, from http://www.cores-eu.net/.

CWI's Semi-automatic Hypermedia Presentation Generation (Dynamo) Project. (2000). Retrieved August 11, 2003, from http://db.cwi.nl/projecten/project.php4?prjnr=74.

DAML+OIL. (2001, December 18). Reference description. W3C Note 18 December 2001. Retrieved August 11, 2003, from http://www.w3.org/TR/daml+oil-reference.

DAML Ontology Library. (2003). Retrieved August 11, 2003, from http://www.daml.org/ontologies/.

DANTE. (2003). Retrieved August 11, 2003, from http://www.dante.net/.

DARPA Object Service Architecture Web Annotation Service (1998). Project Summary. Retrieved August 11, 2003, from http://www.objs.com/OSA/Annotations-Service.html.

Delgado, J., Gallego, I., Garcia, R., & Gil, R. (2002). An ontology for intellectual property rights: IPROnto. Poster at 1st International Semantic Web Conference (ISWC 2002). Retrieved August 11, 2003, from http://dmag.upf.es/flas_eng/publicaciones.htm.

Denny, M. (2002, November). Ontology building: A survey of editing tools. Retrieved August 11, 2003, from http://www.xml.com/pub/a/2002/11/06/ontologies.html.

Department of Computer Science and Engineering, University of Washington (DCSE). (2003). Object and concept recognition for content-based image retrieval. Retrieved August 11, 2003, from http://www.cs.washington.edu/research/imagedatabase/.

De Roure, D., Jennings, N., & Shadbolt, N. (2001). Research agenda for the semantic grid: A future e-science infrastructure. [Technical report]. UKeS-2002–02, UK e-Science Technical Report Series. National e-Science Centre, Edinburgh, UK. Retrieved August 11, 2003, from http://www.semanticgrid.org/html/semgrid.html.

Doctorow, C. (2001). Metacrap: Putting the torch to seven straw-men of the meta-utopia. Retrieved August 11, 2003, from http://www.well.com/~doctorow/metacrap.htm.

Dragon Systems. (2003). Retrieved August 11, 2003, from http://www.dragonsys.com/.
DSpace. (2002). DSpace durable digital depository. Retrieved August 11, 2003, from http://www.dspace.org.
DSTC FilmEd. (2003). The FilmEd project. Retrieved August 11, 2003, from http://metadata.net/filmed/.
DSTC New Media. (2003). The New Media art preservation project. Retrieved August 11, 2003, from http://metadata.net/newmedia/.
Dublin Core Metadata Initiative (DCMI). (2003). Retrieved August 11, 2003, from http://www.dublincore.org/.
Edutella. (2002). Retrieved August 11, 2003, from http://edutella.jxta.org/.
eLib Newsagent project. (1996). Retrieved August 11, 2003, from http://www.ukoln. ac.uk/services/elib/projects/newsagent/.
FAST. (2000). Multimedia Search Engine. Retrieved August 11, 2003, from http://www.multimedia.alltheWeb.com/
Fluhr, C., Schmit, D., Ortet, P., Elkateb, F., & Gurtner, K., (1997). SPIRIT-W3: A distributed cross-lingual indexing and search engine. Retrieved August 11, 2003, from http://www.isoc.org/isoc/whatis/conferences/inet/97/proceedings/A8/A8_1.HTM.
Foaf. (2000). The "Friend of a Friend" Project. Retrieved August 11, 2003, from http://www. foaf-project.org/.
FoafCORP. (2002). Retrieved August 11, 2003, from http://www.grorg.org/2002/10/foafcorp/.
Fossick. (2003). Online multimedia and digital image search. Retrieved August 11, 2003, from http://fossick.com/Multimedia.htm.
Friskit. (2002). Music streaming media search. Retrieved August 11, 2003, from http://www. friskit.com/.
GGF. (2003). Grid initiatives and projects. Retrieved August 11, 2003, from http://www.gridforum.org/L_Involved_Mktg/init.htm.
Gonçalves M. A., France R. K., & Fox, E. A. (2001). MARIAN: Flexible interoperability for federated digital libraries. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2001).* Darmstadt, Germany. Retrieved August 11, 2003, from http://link.springer.de/link/service/series/0558/papers/2163/21630173.pdf.
Google. (2003). Image search. Retrieved August 11, 2003, from http://images.google.com/.
GrangeNet. (2003). Retrieved August 11, 2003, from http://www.grangenet.net/.
Grid Computing. (2000). Retrieved August 11, 2003, from http://www.gridcomputing.com/.
High Performance Wireless Research and Education Network (HPWREN). (2001). Retrieved August 11, 2003, from http://hpwren.ucsd.edu/news/011109.html.
Hunter, J. (2002). Rights markup extensions for the protection of indigenous knowledge. WWW2002 Conference. Honolulu, HI. Retrieved August 11, 2003, from http://archive.dstc.edu.au/IRM_project/paper.pdf.
Hunter, J., Koopman, B., & Sledge, J. (2003). Software tools for indigenous knowledge management. In *Museums on the Web.* Charlotte. Retrieved August 11, 2003, from http://archive.dstc.edu.au/IRM_project/software_paper/IKM_software.pdf.
IBM VideoAnnEx. (2001). Retrieved August 11, 2003, from http://www.research.ibm.com/VideoAnnEx/.
IMS. (2003). IMS Learning Resource Meta-data Specification. Retrieved August 11, 2003, from http://www.imsglobal.org/metadata/index.cfm.
indecs Framework Ltd. (2000). Retrieved August 11, 2003, from http://www.indecs.org/.
Informedia. (2003). Digital video understanding. Retrieved August 11, 2003, from http://www.informedia.cs.cmu.edu/.
Internet2. (2003). Retrieved August 11, 2003, from http://www.internet2.edu/.
IPTC. (2001). NewsML—Markup for the third millenium. Retrieved August 11, 2003, from http://www.iptc.org/site/NewsML/.
Joseph, S. (2003). Decentralized meta-data strategies. University of Tokyo. Retrieved August 11, 2003, from http://www.neurogrid.net/Decentralized_Meta-Data_Strategies-neat.html.
JXTA. (2003). Retrieved August 11, 2003, from http://www.jxta.org.
Kaine-Krolak, M., & Novak, M. (1995). An introduction to infrared technology: Applications in the home, classroom, workplace, and beyond. . . . Retrieved August 11, 2003, from http://trace.wisc.edu/docs/ir_intro/ir_intro.htm.
Kartoo. (2000). Retrieved August 11, 2003, from http://www.kartoo.com/.

Lagoze, C., & Hunter, J. (2001). The ABC ontology and model. *Journal of Digital Information, 2*(2). Retrieved August 11, 2003, from http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/.

Lewis, N. (2002). Talking about a revolution? First impressions of Ex Libris's MetaLib. *Ariadne, 32*. Retrieved August 11, 2003, from http://www.ariadne.ac.uk/issue32/metalib/.

Library of Congress. (2002). Library of Congress: American Memory Historical Collections for the National Digital Library. Retrieved August 11, 2003, from http://memory.loc.gov/.

Library of Congress. (2003). METS (Metadata Encoding and Transmission Standard). Retrieved August 11, 2003, from http://www.loc.gov/standards/mets/.

Little, S., Guerts, J., & Hunter, J. (2002). The dynamic generation of intelligent multimedia presentations through semantic inferencing. ECDL 2002. Rome, Italy. Retrieved August 11, 2003, from http://archive.dstc.edu.au/maenad/ecdl2002/ecdl2002.html.

Liu X., et. al., (2002). Federated searching interface techniques for heterogeneous OAI repositories. *Journal of Digital Information, 2*(4). Retrieved August 11, 2003, from http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/.

Lynch, C. (2001a). Personalization and recommender systems in the larger context: New directions and research questions. Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries, Dublin, Ireland. Retrieved August 11, 2003, from http://www.ercim.org/publication/ws-proceedings/DelNoe02/CliffordLynchAbstract.pdf.

Lynch, C. (2001b). When documents deceive: Trust and provenance as new factors for information retrieval in a tangled Web. *Journal of the American Society for Information Science, 52*(1), 12–17. Retrieved August 11, 2003, from http://www.cs.ucsd.edu/~rik/others/lynch-trust-jasis00.pdf.

Magkanaraki, A., Karvounarakis, G., Anh, T. T., Christophides, V., & Plexousakis, D. (2002). Ontology storage and querying. Technical Report No. 308, ICS FORTH, Crete. Retrieved August 11, 2003, from http://139.91.183.30:9090/RDF/publications/tr308.pdf.

Martinez, J. (2002). MPEG-7 overview (Version 8). Retrieved August 11, 2003, from http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

MSNBC News Tools Home (2003). Retrieved August 11, 2003, from http://www.msnbc.com/toolkit.asp.

Nardi, B. A., & O'Day, V. L. (1998). Application and implications of agent technology for libraries. *The Electronic Library, 16*(5), 325–337.

Net2one Personalized News Informer. (2003). Retrieved August 11, 2003, from http://www.net2one.com/index2.asp

Networked European Deposits Library (NEDLIB). (2001). Retrieved August 11, 2003, from http://www.kb.nl/coop/nedlib/.

Ng, D., Wactlar, H., Hauptmann, A., & Christel, M. (2003). Collages as dynamic summaries of mined video content for intelligent multimedia knowledge management. AAAI Spring Symposium Series on Intelligent Multimedia Knowledge Management. Palo Alto, CA. Retrieved August 11, 2003, from http://www-2.cs.cmu.edu/~hdw/aaai03_ng.pdf.

Oard, D. (2003). Speech retrieval papers and project descriptions. Retrieved August 11, 2003, from http://raven.umd.edu/dlrg/speech/papers.html.

OASIS. (2003). Universal Description, Discovery & Integration (UDDI) of Web Services. Retrieved August 11, 2003, from http://www.uddi.org/.

OCLC/RLG. (2003). Preservation Metadata Working Group. Retrieved August 11, 2003, from http://www.oclc.org/research/pmwg/.

ODRL. (2003). Retrieved August 11, 2003, from http://www.odrl.net/.

Open Archival Information System (OAIS) Resources. (2002). Retrieved August 11, 2003, from http://www.rlg.org/longterm/oais.html.

Open Archives Initiative (OAI). (2003). Retrieved August 11, 2003, from http://www.openarchives.org/.

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). (2003). Version 2.0, June 14, 2002. Retrieved August 11, 2003, from http://www.openarchives.org/OAI/openarchivesprotocol.html.

Open Archives Initiative (OAI) Registered Service Providers. (2002). Retrieved August 11, 2003, from http://www.openarchives.org/service/listproviders.html.

OpenGALEN. (2002). Retrieved August 11, 2003, from http://www.opengalen.org/.
OpenVideo. (2002). The OpenVideo project. Retrieved August 11, 2003, from http://www.open-video.org/.
PANDORA. (2002). National Library of Australia, PANDORA project. Retrieved August 11, 2003, from http://pandora.nla.gov.au/.
PAXit. (2003). PAXit image database software. Retrieved August 11, 2003, from http://www.paxit.com/paxit/communications.asp.
Pretty Good Privacy (PGP). (2002). Retrieved August 11, 2003, from http://www.rubin.ch/pgp/pgp.en.html.
QBIC. (2001). IBM's query by image content. Retrieved August 11, 2003, from http://wwwqbic.almaden.ibm.com/.
Reamy, T., (2002). Auto-categorization: Coming to a library or intranet near you! *EContent Magazine.* Retrieved August 11, 2003, from http://www.econtentmag.com/r5/2002/reamy11_02.html.
Reynolds, D., Cayzer, S., Dickinson, I., & Shabajee, P. (2002). Blogging and semantic blogging. SWAD-Europe Deliverable12.1.1: Semantic Web applications—analysis and selection. Retrieved August 11, 2003, from http://www.w3.org/2001/sw/Europe/reports/chosen_demos_rationale_report/hp-applications-selection.html#sec-appendix-blogging.
Ricoh MovieTool. (2002). Retrieved August 11, 2003, from http://www.ricoh.co.jp/src/multimedia/MovieTool/.
ROADNet. (2002). Real-time observatories, applications, and data management network. Retrieved August 11, 2003, from http://roadnet.ucsd.edu/.
Rogers, D., Hunter J., & Kosovic, D. (2002). The TV-trawler project. *International Journal of Imaging Systems and Technology,* Special Issue on Multimedia Content Description and Video Compression.
RoMEO. (2003). Project RoMEO (Rights MEtadata for Open archiving). Retrieved August 11, 2003, from http://www.lboro.ac.uk/departments/ls/disresearch/romeo.
SCHEMAS. (2002). SCHEMAS—Forum for metadata schema implementers. Retrieved August 11, 2003, from http://www.schemas-forum.org/registry/.
Semantic Grid. (2003). Retrieved August 11, 2003, from http://www.semanticgrid.org/.
Singingfish. (2002). Multimedia search. Retrieved August 11, 2003, from http://www.singingfish.com/.
SNOMED CT. The Systemized Nomenclature of Medicine. (2003). Retrieved August 11, 2003, from http://www.snomed.org/.
Sullivan, A. (2002). The blogging revolution. *Wired,* 10.05. Retrieved August 11, 2003, from http://www.wired.com/wired/archive/10.05/mustread.html?pg=2.
SUO. (2002). IEEE P1600.1 Standard Upper Ontology SUO Working Group. Retrieved August 11, 2003, from http://suo.ieee.org/.
Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., & Wenke, D. (2002). OntoEdit: Collaborative ontology engineering for the semantic Web. In *Proceedings of the First International Semantic Web Conference 2002 (ISWC 2002).* Sardinia, Italy. Retrieved August 11, 2003, from http://link.springer.de/link/service/series/0558/papers/2342/23420221.pdf.
TITAN. (2003). A Cross-Language WWW Search Engine. Retrieved August 11, 2003, from http://titan.mcnet.ne.jp/.
Topic Maps. (2000). Retrieved August 11, 2003, from http://www.topicmaps.org/.
TouchGraph GoogleBrowser. (2001). Retrieved August 11, 2003, from http://www.touchgraph.com/TGGoogleBrowser.html.
University of Illinois Library (UIL). (2002). University of Illinois Open Archives Collection. Retrieved August 11, 2003, from http://bolder.grainger.uiuc.edu/uiLibOAIProvider/2.0/oai.asp.
Virage. (2003). Retrieved August 11, 2003, from http://www.virage.com/.
Virginia Tech. (1997a). Digital libraries and software agents. Retrieved August 11, 2003, from http://scholar.lib.vt.edu/digilib/reports/agents.pdf.
Virginia Tech. (1997b). Ontologies and agents in digital libraries. Retrieved August 11, 2003, from http://ei.cs.vt.edu/~cs6604/f97/agents.htm.
W3C Annotea Web Annotation Service. (2001). Retrieved August 11, 2003, from http://annotest.w3.org/.
W3C RDF Syntax and Model Recommendation. (1999). Retrieved August 11, 2003, from http://www.w3.org/TR/REC-rdf-syntax/.

W3C RDF Vocabulary Description Language 1.0. (2003). RDF Schema, W3C working draft. Retrieved August 11, 2003, from http://www.w3.org/TR/rdf-schema/.

W3C semantic web activity. (2002). Retrieved August 11, 2003, from http://www.w3.org/2001/sw/Activity.

W3C Web Ontology Language (OWL). (2003). Guide, version 1.0, W3C working draft. Retrieved August 11, 2003, from http://www.w3.org/TR/owl-guide/.

W3C Web Ontology (WebOnt) Working Group. (2003). Retrieved August 11, 2003, from http://www.w3.org/2001/sw/WebOnt/.

W3C Web Services Activity. (2003). Retrieved August 11, 2003, from http://www.w3.org/2002/ws/.

W3C Extensible Markup Language (XML). (2003). Retrieved August 11, 2003, from http://www.w3.org/XML.

W3C XML Protocol Working Group. (2003). Simple object access protocol (SOAP). Retrieved August 11, 2003, from http://www.w3.org/2000/xp/Group/.

W3C XML Query. (2003). Retrieved August 11, 2003, from http://www.w3.org/XML/Query.

W3C XML Schema Language. (2003). Retrieved August 11, 2003, from http://www.w3.org/XML/Schema.

W3C XML Signature Working Group. (2003). Retrieved August 11, 2003, from http://www.w3.org/Signature/.

Wang, J. Z., & Li, J. (2003). Evaluation strategies for automatic linguistic indexing of pictures. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. Barcelona, Spain.

WebBrain. (2001). Retrieved August 11, 2003, from http://www.Webbrain.com/.

Web Services Description Language (WSDL). (2003). Version 1.2, W3C working draft. Retrieved August 11, 2003, from http://www.w3.org/TR/wsdl12.

XrML. (2003). Retrieved August 11, 2003, from http://www.xrml.org/.

ZGDV VIDETO. (2002). ZGDV video description tool. Retrieved August 11, 2003, from http://www.rostock.igd.fraunhofer.de/ZGDV/Abteilungen/zr2/Produkte/videto/index_html_en.